Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

Accelerated oblique random survival forests

Byron C. Jaeger BJAEGER@WAKEHEALTH.EDU
Department of Biostatistics and Data Science
Wake Forest University School of Medicine
Winston-Salem, NC 27157, USA

Sawyer Welden SWELDEN@WAKEHEALTH.EDU
Department of Biostatistics and Data Science
Wake Forest University School of Medicine
Winston-Salem, NC 27157, USA

Kristin Lenoir KLENOIR@WAKEHEALTH.EDU
Department of Biostatistics and Data Science
Wake Forest University School of Medicine
Winston-Salem, NC 27157, USA

Jaime L Speiser JSPEISER@WAKEHEALTH.EDU
Department of Biostatistics and Data Science
Wake Forest University School of Medicine
Winston-Salem, NC 27157, USA

Matthew Segar MATTHEW.SEGAR@UTSOUTHWESTERN.EDU
Division of Cardiology, Department of Internal Medicine,
University of Texas Southwestern Medical Center, Dallas

Nicholas M. Pajewski NPAJEWSK@WAKEHEALTH.EDU
Department of Biostatistics and Data Science
Wake Forest University School of Medicine
Winston-Salem, NC 27157, USA

Editor: TBD

Abstract

The oblique random survival forest (RSF) is an ensemble supervised learning method for right-censored outcomes. Trees in the oblique RSF are grown using linear combinations of predictors to create branches, whereas in the standard RSF a single predictor is used. Oblique RSF ensembles often have higher prediction accuracy than standard RSF ensembles, but the additional computational overhead of assessing all possible linear combination of predictors is a severe limitation. In addition, few methods have been developed for interpretation of oblique RSF ensembles, and they remain something of a black box compared to their axis-based counterparts. In this article, we introduce and evaluate a method to increase computational efficiency of the oblique RSF and a method to estimate importance of individual predictor variables with the oblique RSF. Our strategy to reduce computational overhead makes use of Newton-Raphson scoring, a classical optimization technique that we apply to the Cox partial likelihood function within each non-leaf node of decision trees. We estimate importance of predictors for the oblique RSF by negating each coefficient used for the given predictor in linear combinations, and then computing the reduction in out-of-bag accuracy. In numeric experiments, we find that our implementation of the oblique RSF is over 500 times faster with equivalent discrimination and superior Brier score compared to existing software for oblique RSFs. We find in simulation studies that 'negation importance' discriminates between signal and noise predictors more reliably than permutation importance, Shapley additive explanations, and a previously introduced technique to measure variable importance with oblique RSFs based on analysis of variance. All methods pertaining to oblique RSFs in the current study are available in the `aorsf` R package.

Keywords: Random Forests, Survival, Efficient, Variable Importance

1. Introduction

Risk prediction may reduce the burden of disease by guiding strategies for prevention and treatment in a wide range of domains (Moons et al., 2012a,b). The random survival forest (RSF; Ishwaran et al. (2008); Hothorn et al. (2006)) is a supervised learning algorithm that has been used frequently for risk prediction (Wang and Li, 2017). Similar to random forests (RFs) for classification and regression (Breiman, 2001), The RSF is a large set of de-correlated and randomized decision trees, with each tree contributing to the ensemble's prediction function. Notable characteristics of the RSF include uniform convergence of its ensemble survival prediction function to the true survival function, first shown by Ishwaran and Kogalur (2010) and later by Cui et al. (2017) under more general conditions. However, Cui et al. (2017) noted that the RSF is at a disadvantage when predictors are correlated and some are not relevant to the censored outcome, which is a strong possibility when large medical databases are leveraged for risk prediction.

A potential approach to improve the RSF when predictors are correlated and some are not relevant to the censored outcome is to use oblique trees instead of axis based trees. Axis based trees split data using a single predictor, creating decision boundaries that are perpendicular or parallel to axes of the predictor space (see Breiman et al., 2017,

Chapter 2). Oblique trees split data using a linear combination of predictors, creating decision boundaries that are neither parallel nor perpendicular to axes of their contributing predictors (see Breiman et al., 2017, Chapter 5). Menze et al. (2011) examined prediction accuracy of RFs in the presence of correlated predictors and found that oblique RFs had substantially higher prediction accuracy compared to axis-based RFs. Similarly, Jaeger et al. (2019) found that growing RSFs with oblique rather than axis-based survival trees reduced the RSF's concordance error, with improvements ranging from 2.5% to 24.9% depending on the data analyzed.

Oblique trees have at least two notable drawbacks compared to axis-based trees. First, finding a locally optimal oblique decision rule may require exponentially more computation than an axis-based rule. If p predictors are potentially used to split n observations, up to $\mathcal{O}(n^p)$ oblique splits can be assessed versus $\mathcal{O}(n \cdot p)$ axis-based splits (Heath et al., 1993; Murthy et al., 1994). Second, estimating variable importance (VI) using permutation (a standard method for RFs) may be less effective in ensembles of oblique trees, as permuting the values of one predictor may not destabilize decisions that are based on many predictors. Although VI is one of the most widely used strategies to interpret random forests (Ishwaran and Lu, 2019), few studies have investigated VI for oblique random forests (see Menze et al., 2011, Section 5), and fewer have investigated VI specifically for the oblique RSF.

This study makes two contributions to oblique RSFs. First, we reduce their computational cost (that is, accelerate them) with a scalable algorithm to identify linear combinations of coefficients. In a general benchmark experiment including 30 risk prediction tasks, we show that accelerated oblique RSFs are roughly 500 times faster with equivalent or superior prediction accuracy compared to existing routines to fit oblique RSFs. Second, we improve the interpretability of oblique RSFs with 'negation VI', a method to estimate VI that flips the sign of coefficients in linear combinations of predictors instead of permuting predictor values. In simulation studies where VI is estimated using permutation, analysis of variance (ANOVA; see Menze et al. (2011)), and approximations of Shapley values, we find that negation VI improves the oblique RSF's ability to discriminate between correlated signal and noise variables. The accelerated oblique RSF and multiple methods to compute VI for oblique RSFs (permutation, ANOVA, and negation) are available in the `aorsf` R Package.

2. Related work

Sections 2.1 and 2.2 briefly summarize prior studies that have developed methods related to the oblique RSF and VI, respectively.

2.1 Axis-based and oblique random forests

After Breiman (2001) introduced the axis-based and oblique RF, numerous methods were developed to grow oblique RFs for classification or regression tasks (Menze et al., 2011;

Zhang and Suganthan, 2014; Rainforth and Wood, 2015; Zhu et al., 2015; Poona et al., 2016; Qiu et al., 2017; Tomita et al., 2020; Katuwal et al., 2020). However, oblique splitting approaches for classification or regression may not generalize to survival (for example, see Zhu, 2013, Section 4.5.1), and most research involving the RSF has focused on forests with axis-based trees (Wang and Li, 2017).

Building on prior research for bagging survival trees (Hothorn et al., 2004), Hothorn et al. (2006) developed an axis-based RSF in their framework for unbiased recursive partitioning, more commonly referred to as the conditional inference forest (CIF). Zhou et al. (2016) developed a rotation forest based on the CIF and Wang and Zhou (2017) developed a method for extending the predictor space of the CIF. Ishwaran et al. (2008) developed an axis-based RSF with strict adherence to the rules for growing trees proposed in Breiman (2001). A similar implementation of the RSF was implemented in the `ranger` (Wright and Ziegler, 2017), an R package designed for high dimensional data that allows users to fit extremely randomized survival forests (Geurts et al., 2006). Jaeger et al. (2019) developed the oblique RSF following the bootstrapping approach described in Breiman's original RF and incorporating early stopping rules from the CIF.

2.2 Variable importance

Breiman (2001) introduced permutation VI, defined for each predictor as the difference in a RF's estimated generalization error before versus after the predictor's values are randomly permuted. Strobl et al. (2007) identified bias in permutation VI driven by variable selection bias and effects induced by bootstrap sampling, and proposed an unbiased permutation VI based on unbiased recursive partitioning (see Hothorn et al. (2006)). Menze et al. (2011) introduced an approach to estimate VI for oblique RFs that computes an analysis of variance (ANOVA) table in non-leaf nodes to obtain p-values for each predictor contributing to the node. The ANOVA VI[1] is then defined for each predictor as the number of times a p-value associated with the predictor is ≤ 0.01 while growing a forest. Lundberg and Lee (2017) introduced a method to estimate VI using SHapley Additive exPlanation (SHAP) values, which estimate the contribution of a predictor to a model's prediction for a given observation. SHAP VI is computed for each predictor by taking the mean absolute value of SHAP values for that predictor across all observations in a given set.

3. The acceleracted oblique random survival forest

Consider the usual framework for survival analysis with training data

$$\mathcal{D}_{\text{train}} = \{(T_i, \delta_i, \boldsymbol{x}_i)\}_{i=1}^{N_{\text{train}}} .$$

1. Menze et al. (2011) name their method 'oblique RF VI', but we use the name 'ANOVA VI' in this article to avoid confusing Menze's approach with other approaches to estimate VI for oblique RFs.

Here, T_i is the event time if $\delta_i = 1$ and last point of contact if $\delta_i = 0$, and $\boldsymbol{x}_i$ is a vector of predictors values. Assuming there are no ties, let $t_1 < \ldots < t_m$ denote the m unique event times in $\mathcal{D}_{\text{train}}$.

To accelerate the oblique RSF, we propose to identify linear combinations of predictor variables in non-leaf nodes by applying Newton Raphson scoring to the partial likelihood function of the Cox regression model:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{e^{\boldsymbol{x}_{j(i)}^T \boldsymbol{\beta}}}{\sum_{j \in R_i} e^{\boldsymbol{x}_j^T \boldsymbol{\beta}}}, \tag{1}$$

where R_i is the set of indices, j, with $T_j \geq t_i$ (i.e., those still at risk at time t_i), and $j(i)$ is the index of the observation for which an event occurred at time t_i. Newton Raphson scoring is an extremely fast estimation procedure, and the `survival` package includes documentation that outlines how to efficiently program it (Therneau, 2022). Briefly, a vector of estimated regression coefficients, $\hat{\beta}$, is updated in each step of the procedure based on its first derivative, $U(\hat{\beta})$, and second derivative, $H(\hat{\beta})$:

$$\hat{\beta}^{k+1} = \hat{\beta}^k + U(\hat{\beta} = \hat{\beta}^k)\, H^{-1}(\hat{\beta} = \hat{\beta}^k)$$

For statistical inference, it is recommended to complete iterations of Newton Raphson scoring until a convergence threshold is met. However, to identify a valid linear combination of predictors, only one iteration of Newton Raphson scoring is needed. In Section 4.1.6, we formally test whether growing oblique survival trees using one iteration of Newton Raphson scoring is sufficient (that is, provides equivalent prediction accuracy) compared to iterating until a convergence threshold is met.

Algorithm 1 presents our approach to fitting an oblique survival tree in the accelerated oblique RSF using default values from the `aorsf` R package. Several steps are taken to reduce computational overhead. First, memory is conserved by conducting bootstrap resampling via randomly generated bootstrap weights. Weights are integer valued, with a weight of v indicating an observation was sampled v times. Second, early stopping is applied to the tree growing procedure if a statistical criterion is not met. In our case, the criterion is based on the magnitude of a log-rank test statistic corresponding to splitting the data at a current node. Third, instead of greedy recursive partitioning, we use a 'good enough' approach. More specifically, instead of computing a log-rank test statistic for several different linear combinations of variables and proceeding with the highest scoring option, we identify an optimal cut-point for one linear combination of variables and assess whether using this combination will create a split that passes the criterion for splitting a node. If it does not pass the criterion, then another linear combination will be tested, with the maximum number of attempts set by the parameter `n_retry`. Often a 'good-enough' split can be found in just one attempt when the training set is large, which gives the accelerated oblique RSF a computational advantage in larger training sets compared to greedy partitioning.

Algorithm 1 Accelerated oblique random survival tree using default parameters.

Require: Training data $\mathcal{D}_{\text{train}} = \{(T_i, \delta_i, \boldsymbol{x}_i)\}_{i=1}^{N_{\text{train}}}$, mtry $= \sqrt{\text{ncol}(\boldsymbol{x}_{\text{train}})}$, n_split $= 5$, n_retry $= 3$, and split_min_stat $= 3.841459$

$\mathcal{T} \leftarrow \emptyset$
$w \leftarrow$ sample(from $= \{0, \ldots, 10\}$, size $=$ nrow($\boldsymbol{x}_{\text{train}}$), replace $=$ T)
$\mathcal{D}_{\text{in-bag}} \leftarrow$ subset($\mathcal{D}_{\text{train}}$, rows $=$ which($w > 0$))
$w \leftarrow$ subset(w, which($w > 0$))
node_assignments $\leftarrow$ rep(1, times $=$ nrow($\boldsymbol{x}_{\text{in-bag}}$))
nodes_to_split $\leftarrow \{1\}$
repeat
  for node $\in$ nodes_to_split **do**
    n_try $\leftarrow 1$
    node_rows $\leftarrow$ which(node_assignments $\equiv$ node)
    node_cols $\leftarrow$ sample(from $= \{1, \ldots, \text{ncol}(\boldsymbol{x})\}$, size $=$ mtry, replace $=$ F)
    $\mathcal{D}_{\text{node}} \leftarrow$ subset($\mathcal{D}_{\text{in-bag}}$, rows $=$ node_rows, columns $=$ node_cols)
    $\beta \leftarrow$ newt_raph($\mathcal{D}_{\text{node}}$, weights $=$ subset(w, node_rows), max_iter $= 1$)
    $\eta \leftarrow \boldsymbol{x}_{\text{node}} \times \beta$
    $\mathcal{C} \leftarrow$ sample(from $=$ unique(η), size $=$ n_split, replace $=$ F)
    $c \leftarrow \text{argmax}_{c^* \in \mathcal{C}} \{\text{log\_rank\_stat}(\eta, c^*)\}$
    if log_rank_stat$(\eta, c) \geq$ split_min_stat **then**
      $\mathcal{T} \leftarrow$ add_node($\mathcal{T}$, name $=$ node, beta $= \beta$, cutpoint $= c$)
      ▷ Right node logic omitted for brevity (identical to left node logic)
      node_left_name $\leftarrow$ max(node_assignments) $+ 1$
      node_left_rows $\leftarrow$ subset(node_rows, which($\eta \leq c$))
      subset(node_assignments, node_left_rows) $\leftarrow$ node_left_name
      if is_splittable(subset(node_assignments, node_left_rows)) **then**
        nodes_to_split $\leftarrow$ nodes_to_split $\cup$ node_left_name
      else
        $\mathcal{T} \leftarrow$ add_leaf($\mathcal{T}$, data $=$ subset($\mathcal{D}_{\text{node}}$, rows $=$ node_left_rows))
      end if
    else if n_try $\leq$ n_retry **then**
      n_try $\leftarrow$ n_try $+ 1$
      go to 11
    else
      $\mathcal{T} \leftarrow$ add_leaf($\mathcal{T}$, data $= \mathcal{D}_{\text{node}}$)
    end if
    nodes_to_split $\leftarrow$ nodes_to_split $\setminus$ node
  end for
until nodes_to_split $= \emptyset$
return $\mathcal{T}$

3.1 Negation variable importance

Negation VI is similar to permutation VI in that it measures how much a model's prediction error increases when a variable's role in the model is de-stabilized. Specifically, negation VI measures the increase in an oblique RF's prediction error after flipping the sign of all coefficients linked to a variable (that is, negating them). As the magnitude of a coefficient increases, so does the probability that negating it will change the oblique RF's predictions. For the current study, we use Harrell's concordance (C)-statistic (Harrell et al., 1982) to measure change in prediction error when computing negation VI.

Negation VI has several helpful characteristics. First, negation VI can be applied to any oblique RF using any valid error function.[2] Second, since the coefficients in each non-leaf node of an oblique tree are adjusted for the accompanying predictors, negation VI may provide better estimation of VI in the presence of correlated variables compared to standard VI techniques. Third, unlike permutation, negation is non-random and hence reproducible without setting a random seed. Additionally, since negation VI does not permute variables, the analyst need not worry about impossible combinations of predictors that may occur when one predictor is randomly permuted, such as having a negative status for type 2 diabetes and having Hemoglobin A1c level $\geq 6.5\%$ as a result of randomly permuting the values of Hemoglobin A1c.

4. Numeric experiments

Sections 4.1 and 4.2 present numerical experiments examining the accelerated oblique RSF and negation VI, respectively. The code used to run these experiments is available online at https://github.com/bcjaeger/aorsf-bench. All analyses were conducted using R version 4.1.3 with assistance from multiple R packages. To standardize comparisons of computational efficiency, all learners and VI techniques used up to 4 processing units.

4.1 Benchmark of prediction accuracy and computational efficiency

The aim of this numeric experiment is to evaluate and compare the accelerated oblique RSF with its predecessor (the oblique RSF from the `obliqueRSF` R package) and with other machine learning algorithms for risk prediction. Inferences drawn from this experiment include equivalence and inferiority tests based on Bayesian linear mixed models.

4.1.1 Learners

We consider four classes of learners: RSFs (both axis based and oblique), boosting ensembles, regression models, and neural networks (Table 1). For each class, we synchronized shared tuning parameters. For example, for RSF learners, we set the minimum node size (a

2. The `aorsf` package enables customized functions to be applied in lieu of the default C-statistic (see `?aorsf::orsf_vi_negate`)

parameter shared by all RSF learners) as 10. Additionally, for RSF learners, the number of randomly selected predictors was the square root of the total number of predictors rounded to the nearest integer, and the number of trees in the ensemble was 500. For boosting, regression, and neural network learners, nested 10-fold cross-validation was applied to tune relevant model parameters. Specifically, tuning for boosting models included identifying the number of steps to complete. For regression models, tuning was used to identify the magnitude of penalization. For neural networks, the number and density of layers was tuned.

Learner Class	Software	Learners	Description
Random Survival Forests			

Axis based	`RandomForestSRC`
	`ranger`
	`party`
	`rotsf`
	`rsfse`
		`rsf-standard`
		`rsf-extratrees`
		`cif-standard`
		`cif-rotate`
		`cif-spacextend`
			`rsf-standard` grows survival trees following Leo Breiman's original random forest algorithm with variables and cut-points selected to maximize a log-rank statistic. `rsf-extratrees` grows survival trees with randomly selected features and cut-points. `cif-standard` uses the framework of conditional inference to grow survival trees. `cif-rotate` extends `cif-standard` by applying principal component analysis to random subsets of data prior to growing each survival tree. `cif-spacextend` derives new predictors for each tree in the ensemble, separately.

Oblique	`obliqueRSF`
	`aorsf`
		`obliqueRSF-net`
		`aorsf-net`
		`aorsf-fast`
		`aorsf-cph`
		`aorsf-extratrees`
			Oblique survival trees following Leo Breiman's random forest algorithm. Linear combinations of inputs are derived using `glmnet` in `obliqueRSF-net` and `aorsf-net`, using Newton Raphson scoring for the Cox partial likelihood function in `aorsf-fast` (1 iteration of scoring) and `aorsf-cph` (up to 20 iterations), and chosen randomly from a uniform distribution in `aorsf-extratrees`. Cut-points are selected from 5 randomly selected candidates to maximize a log-rank statistic.

Boosting ensembles			
Trees	`xgboost`	`xgboost-cox` `xgboost-aft`	`xgboost-cox` maximizes the Cox partial likelihood function, whereas `xgboost-aft` maximizes the accelerated failure time likelihood function. Nested cross validation (5 folds) is applied to tune the number of trees grown, the minimum number of observations in a leaf node was 10, the maximum depth of trees was 6, and $\sqrt{p}$ variables were considered randomly for each tree split, where p is the total number of predictors.
Regression models			
Cox Net	`glmnet`	`glmnet-cox`	The Cox proportional hazards model is fit using an elastic net penalty. Nested cross validation (5 folds) is applied to tune penalty terms.
Neural networks			
Cox Time	`survivalmodels`	`nn-cox`	A neural network based on the proportional hazards model with time-varying effects. Nested cross-validation was applied to select the number of layers (from 1 to 8), the number of nodes in each layer (from $\sqrt{p}/2$ to $\sqrt{p}$), and the number of epochs to complete (up to 500). A drop-out rate of 10% was applied during training.

Table 1: Learning algorithms assessed in numeric studies. `aorsf-fast` is the accelerated oblique random survival forest (see Algorithm 1), and each of the additional learners are compared to `aorsf-fast` in numeric studies.

4.1.2 Evaluation of prediction accuracy

Our primary metric for evaluating the accuracy of predicted risk is the integrated and scaled Brier score (Graf et al., 1999). Consider a testing data set:

$$\mathcal{D}_{\text{test}} = \{(T_i, \delta_i, x_i)\}_{i=1}^{N_{\text{test}}}.$$

Let $\widehat{S}(t \mid x_i)$ be the predicted probability of survival up to a given prediction horizon of $t > 0$. For observation i in $\mathcal{D}_{\text{test}}$, let $\widehat{S}(t \mid \boldsymbol{x}_i)$ be the predicted probability of survival up to a given prediction horizon of $t > 0$. Define

$$\widehat{\text{BS}}(t) = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \{\widehat{S}(t \mid \boldsymbol{x}_i)^2 \cdot I(T_i \leq t, \delta_i = 1) \cdot \widehat{G}(T_i)^{-1} + [1 - \widehat{S}(t \mid \boldsymbol{x}_i)]^2 \cdot I(T_i > t) \cdot \widehat{G}(t)^{-1}\}$$

where $\widehat{G}(t)$ is the Kaplan-Meier estimate of the censoring distribution. As $\widehat{\text{BS}}(t)$ is time dependent, integration over time provides a summary measure of performance over a range of plausible prediction horizons. The integrated $\widehat{\text{BS}}(t)$ is defined as

$$\widehat{\mathcal{BS}}(t_1, t_2) = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \widehat{\text{BS}}(t) dt. \tag{2}$$

In our results, t_1 and t_2 are the 25th and 75th percentile of event times, respectively. $\widehat{\mathcal{BS}}(t_1, t_2)$, a sum of squared prediction errors, can be scaled to produce a measure of explained residual variation (that is, an R^2 statistic) by computing

$$R^2 = 1 - \frac{\widehat{\mathcal{BS}}(t_1, t_2)}{\widehat{\mathcal{BS}}_0(t_1, t_2)} \tag{3}$$

where $\widehat{\mathcal{BS}}_0(t_1, t_2)$ is the integrated Brier score when a Kaplan-Meier estimate for survival based on the training data is used as the survival prediction function $\widehat{S}(t)$. We refer to this R^2 statistic as the index of prediction accuracy (IPA) (Kattan and Gerds, 2018).

Our secondary metric for evaluating predicted risk is the time-dependent concordance (C)-statistic. We compute the first time-dependent C-statistic proposed by Blanche et al. (2013, Equation 3), which is interpreted as the probability that a risk prediction model will assign higher risk to a case (that is, an observation with $T \leq t$ and $\delta = 1$) versus a non-case (that is, an observation with $T > t$). Similar to the IPA, observations with $T \leq t$ and $\delta = 0$ only contribute to inverse proportion of censoring weights for the time-dependent C-statistic.

Both the IPA and time-dependent C-statistic generally take values between 0 and 1. To avoid presenting an excessive amount of leading zeroes in our tables, figures, and text, we scale both the IPA and time-dependent C-statistic by 100. For example, we present a value of 25 if the IPA is 0.25, 87 if the time-dependent C-statistic is 0.87, and present 10.2 if the difference between two IPA values is 0.102

4.1.3 Data sets

We use a collection of 17 data sets to benchmark the prediction accuracy and computational efficiency of the accelerated ORSF and each of the other learners described in Section 4.1.1. The number of right-censored outcomes per data set ranged from one to four, and the total number of risk prediction tasks we analyzed was 30 (Table A.1). Across all prediction tasks, the number of observations ranged from 137 to 17,549 (median: 1,807.5), the number of predictors ranged from 7 to 1,692 (median: 32.5), and the percentage of censored observations ranged from 5.26 to 97.7 (median: 78.1).

4.1.4 Monte-Carlo cross validation

For each risk prediction task, we completed 22 runs of Monte-Carlo cross validation. In each run, we used a random sample containing 50% of the available data for training and the remaining 50% for testing each of the learners described in Section 4.1.1. Then, for each learner, we computed the IPA, time-dependent C-statistic, and computational time required to fit a prediction model and compute risk predictions. If any learner failed to obtain predictions on any particular split of data[3], the results for that split were omitted from downstream analyses.

4.1.5 Statistical analysis

After collecting data from 22 replications of Monte-Carlo cross validation for the 14 learners in all 30 risk prediction tasks, we analyzed the resulting 9,240 observations of IPA and, separately, time-dependent C-statistic, using a Bayesian linear mixed model. Our approach follows the ideas described by Benavoli et al. (2017) and Kuhn and Wickham (2020), who developed guidelines on making statistical comparisons between learners using Bayesian models. Specifically, we fit two models:

$$\text{IPA} = \widehat{\gamma}_0 + \widehat{\gamma} \cdot \text{learner} + (1 \,|\, \text{data/run})$$

and

$$\text{C-stat} = \widehat{\gamma}_0 + \widehat{\gamma} \cdot \text{learner} + (1 \,|\, \text{data/run}).$$

Random intercepts for specific splits of data (that is, `run` in the model formula) were nested within datasets. The intercept, $\widehat{\gamma}_0$, was the expected value of the outcome using `aorsf-fast`, making the coefficients in $\widehat{\gamma}$ the expected differences between `aorsf-fast` and other learners. Default priors from `rstanarm` were applied for model fitting (Goodrich et al., 2022).

3. For example, when the prediction task was to predict risk of death in the ACTG 320 clinical trial (26 events total), some splits did not leave enough events in the training data to fit complex learners such as the neural network

Hypothesis testing For both the IPA and time-dependent C-statistic, we conducted equivalence and inferiority tests based on a 1 point region of practical equivalence. More specifically, we concluded that two learners had practically equivalent IPA or time-dependent C-statistic if there was a 95% or higher posterior probability that the absolute difference in the relevant metric was less than 1. We concluded that one learner was weakly superior when there was $\geq 95\%$ posterior probability that the difference in the relevant metric was non-zero, and concluded superiority when when there was $\geq 95\%$ posterior probability that the difference in the relevant metric was 1 or more.

4.1.6 Results

A full summary of all results presented in this Section is provided in Table A.2.

Index of prediction accuracy Compared to learners that were not oblique RSFs, `aorsf-fast` had the highest IPA in 16 out of 30 risk prediction tasks, with an overall mean IPA of 13.6 (Figure 1). Compared to the learner with the second highest mean IPA (`rsf-standard`), `aorsf-fast`'s mean was 1.32 points higher, a relative increase of 10.7%. The posterior probability of `aorsf-fast` and `aorsf-cph` having practically equivalent expected IPA was 0.97, and the posterior probability of `aorsf-fast` having a superior IPA to other learners ranged from 0.79 (versus `rsf-standard`) to >0.999 (versus several other learners; see Figure 2)

Time-dependent concordance statistic Compared to learners that were not oblique RSFs, `aorsf-fast` had the highest time-dependent C-statistic in 10 out of 30 risk prediction tasks, with an overall mean of 77.6 (Figure 3). Compared to the learner with the second highest mean C-statistic (`cif-standard`), `aorsf-fast`'s mean was 0.664 points higher, a relative increase of 0.864%. The posterior probability of `aorsf-fast` and `aorsf-cph` having practically equivalent expected time-dependent C-statistics was > 0.999, and the posterior probability of `aorsf-fast` having a superior time-dependent C-statistic versus other learners ranged from 0.19 (versus `cif-standard`) to >0.999 (versus several other learners; see Figure 4)

Computational efficiency Overall, `aorsf-fast` was the second fastest learner, with an expected model development and risk prediction time about 1/2 second longer than `glmnet-cox` (Figure 5). Comparing median computing times, `aorsf-fast` was 525.2 times faster than its predecessor, `obliqueRSF-net`. In addition, `aorsf-fast` was 20.5, 4.27, and 2.79 faster than axis based forests grown using the `party`, `ranger`, and `randomForestSRC` packages, respectively.

4.2 Benchmark of variable importance

The aim of this experiment is to evaluate negation VI and similar VI methods based on how well they can discriminate between variables that do or do not have a relationship

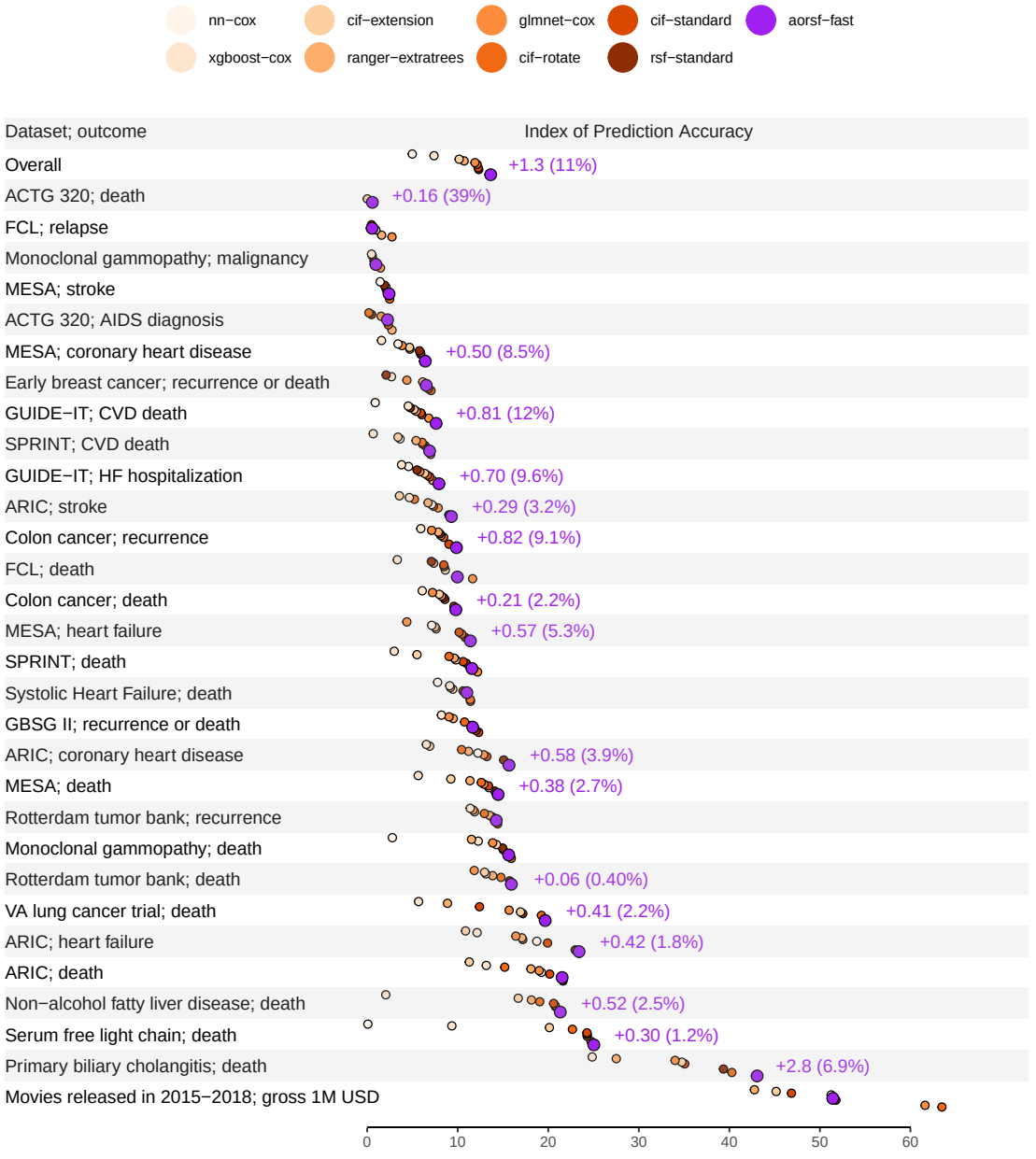

Figure 1: Index of prediction accuracy for the accelerated oblique random survival forest and other learning algorithms across multiple risk prediction tasks. Text appears in tasks where the accelerated oblique random survival forest obtained the highest index of prediction accuracy, showing the absolute and percent improvement over the second best learner. As predicted survival probabilities are not a standard output from `xgboost-aft`, it is not included in this figure.

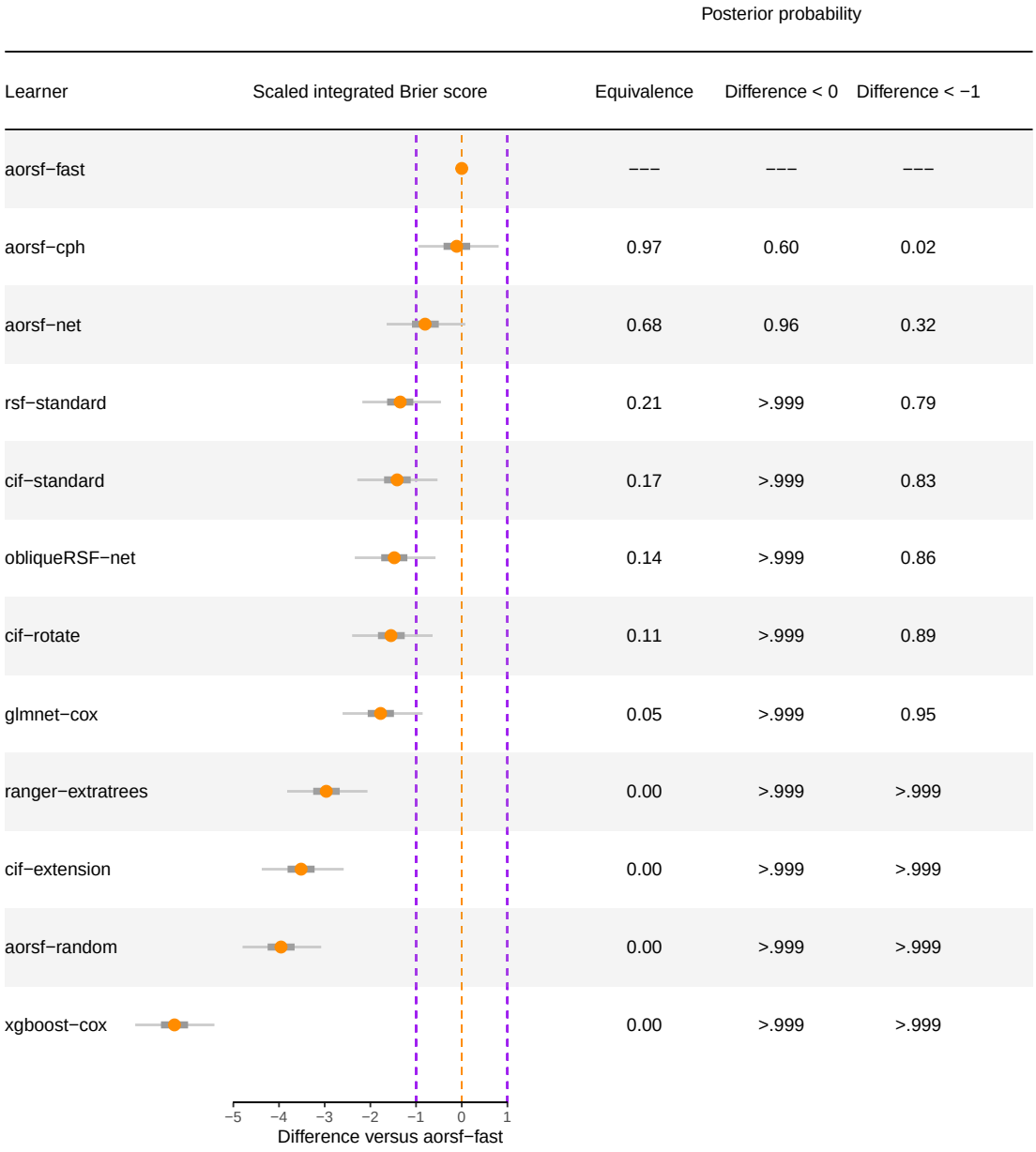

Figure 2: Expected differences in index of prediction accuracy between the accelerated oblique random survival forest and other learning algorithms. A region of practical equivalence is shown by purple dotted lines, and a boundary of non-zero difference is shown by an orange dotted line at the origin.

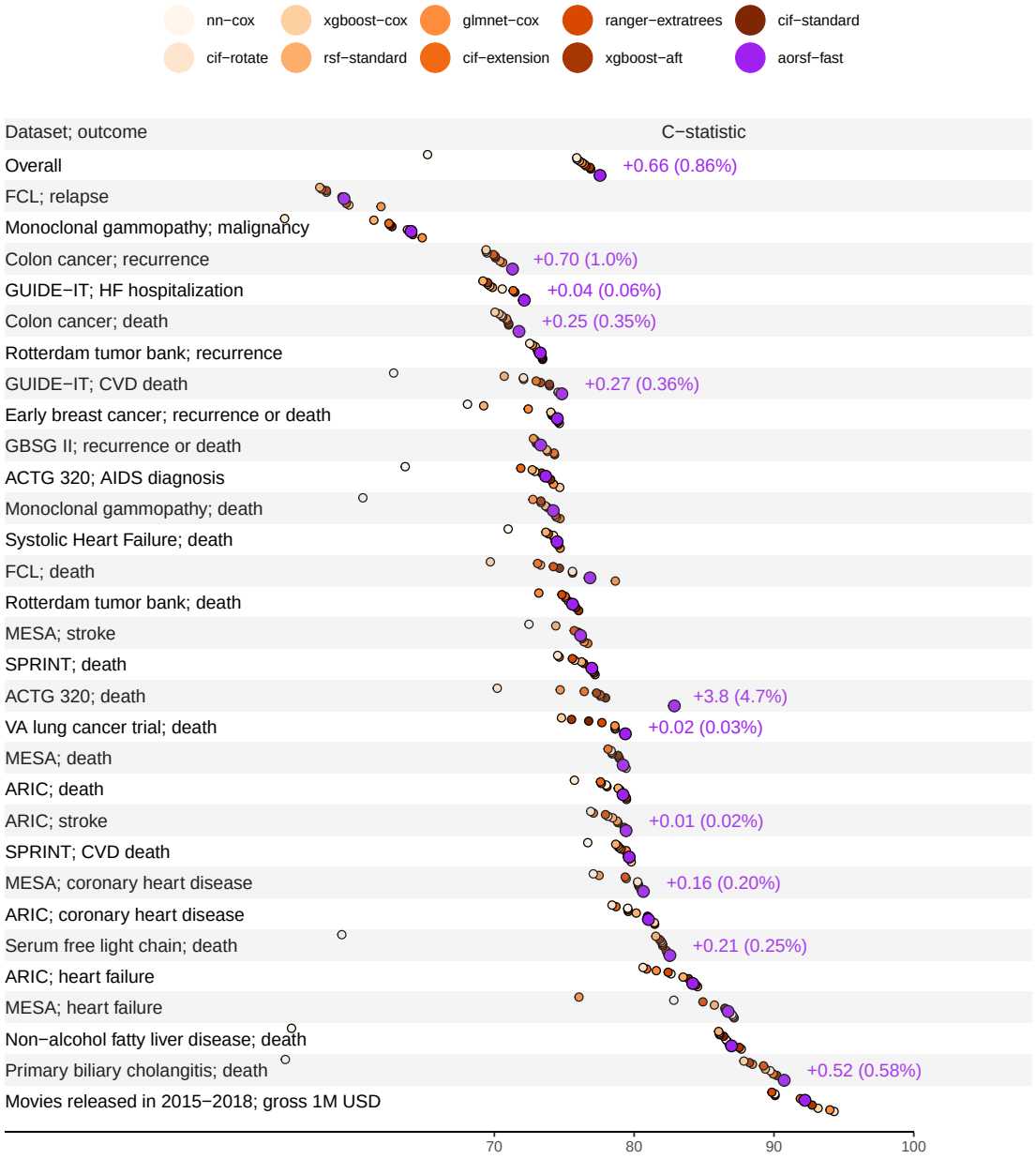

Figure 3: Time-dependent concordance statistic for the accelerated oblique random survival forest and other learning algorithms across multiple risk prediction tasks. Text appears in tasks where the accelerated oblique random survival forest obtained the highest concordance, showing the absolute and percent improvement over the second best learner.

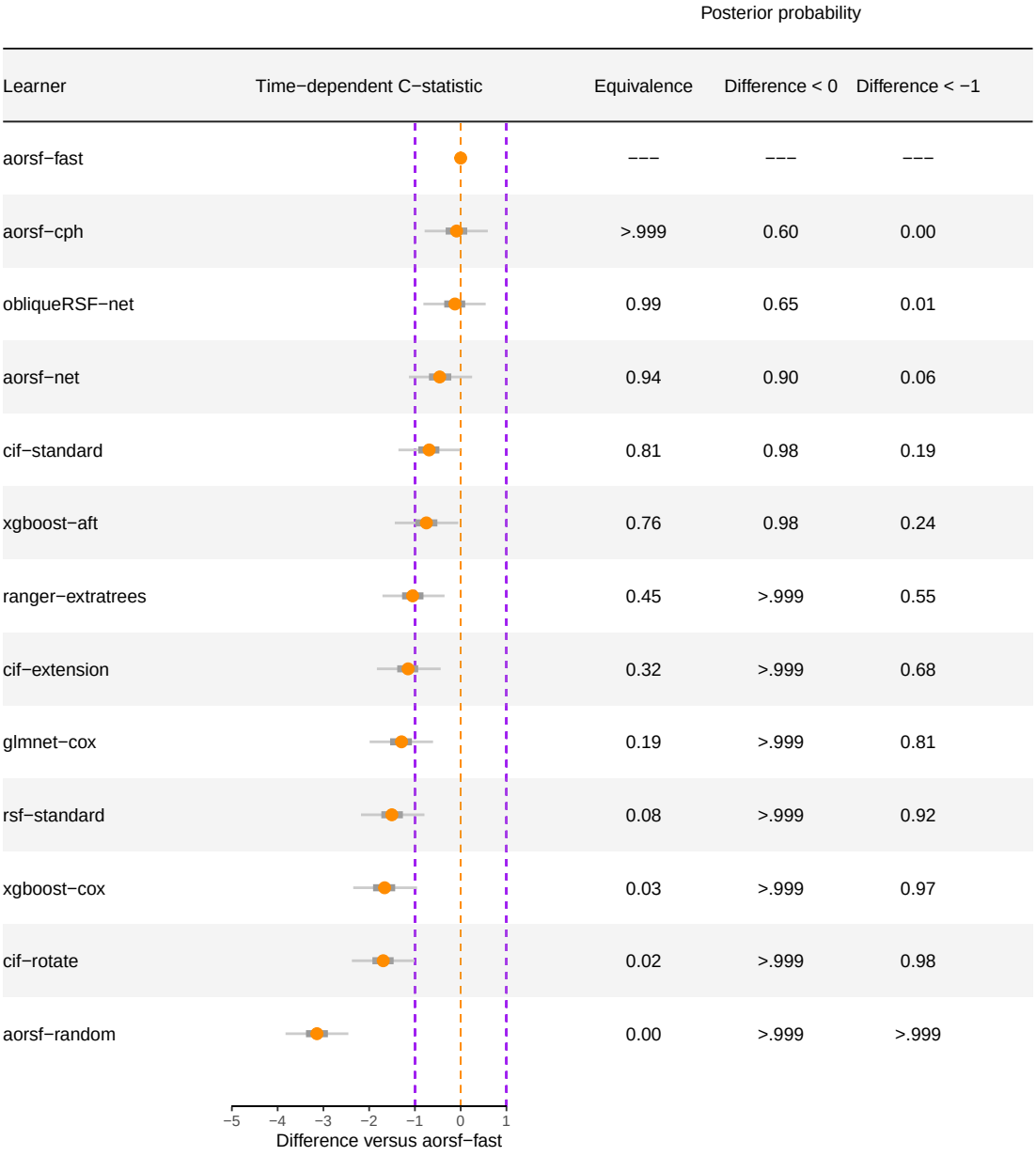

Figure 4: Expected differences in time-dependent concordance statistic between the accelerated oblique random survival forest and other learning algorithms. A region of practical equivalence is shown by purple dotted lines, and a boundary of non-zero difference is shown by an orange dotted line at the origin.

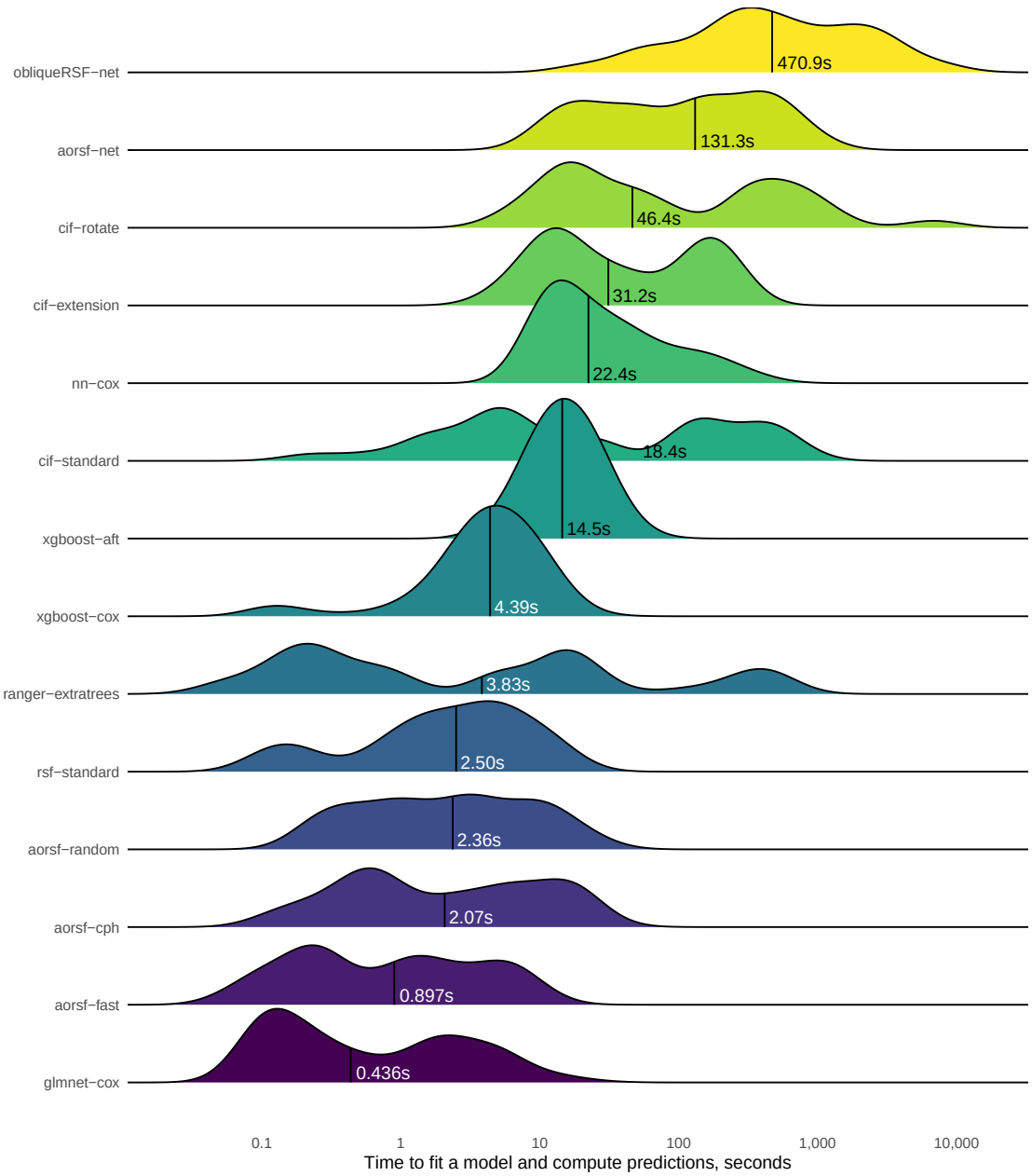

Figure 5: Distribution of time taken to fit a prediction model and compute predicted risk. The median time, in seconds, is printed and annotated for each learner by a vertical line.

with a simulated outcome. We consider methods that are intrinsic to the oblique RF (for example, ANOVA VI), those that are intrinsic to the RF (for example, permutation VI), and those that are model-agnostic (for example, SHAP VI). VI methods with unavailable or still developing software were not included.[4]

4.2.1 Variable importance techniques

We compute permutation VI for axis based RSFs using the `randomForestSRC` package. We compute ANOVA VI, negation VI, and permutation VI for oblique RSFs using the `aorsf` package. For ANOVA VI, we applied a p-value threshold of 0.01, following the threshold recommended by Menze et al. (2011). We compute SHAP VI for boosted tree models using the `xgboost` package, which incorporates the tree SHAP approach proposed by Lundberg et al. (2018).

4.2.2 Variable types

We considered five classes of predictor variables, with each class characterized by its variables' relationship to a right-censored outcome. Specifically,

- *irrelevant* variables had no relationship with the outcome.
- *main effect* variables had a linear relationship to the outcome.
- *non-linear effect* variables had a non-linear relationship to the outcome.
- *combination effect* variables were formed by linear combinations of three other variables. While their combination was linearly related to the outcome, each of the three variables contributing to the combination had no relation to the outcome.
- *interaction effect* variables were related to the outcome by multiplicative interaction with one other variable, which could have been a main effect, non-linear effect, or combination effect variable.

4.2.3 Simulated data

We initiated each set of simulated data with a random draw of size n from a p-dimensional multivariate normal distribution, yielding n observations of p predictors. Each of p predictor variables had a mean of zero, standard deviation of 1, and correlation with other predictor variables drawn at random between a lower and upper boundary. A time-to-event outcome with roughly 45% of observations censored was generated using the `simsurv` package. The full predictor matrix (that is, including interactions, non-linear mappings, and

4. Although the `party` package implements the approach to VI developed by Strobl et al. (2007), the developers of the `party` package note that the implementation of this approach for survival outcomes is "extremely slow and experimental" as of version 1.3.10. Therefore, it is not incorporated in the current simulation study.

combinations) was used to generate the outcome. Interactions, non-linear mappings, and combinations were dropped from the predictor matrix after the outcome was generated so that VI techniques could be evaluated based on their ability to detect these effects.

4.2.4 Parameter specifications

Parameters that varied in the current simulation study included the number of observations (500, 1000, and 2500) and the absolute value of the maximum correlation between predictors (0.3, 0.15, and 0). Parameters that remain fixed throughout the study included the number of predictors in each class (15) and the effect size of each predictor (one standard deviation increase associated with a 64% increase in relative risk).

4.2.5 Evaluation of variable importance

We compared VI techniques based on their discrimination (that is, C-statistic) between relevant and irrelevant variables. Specifically, we generated a binary outcome for each predictor variable based on its relevance (that is, the binary outcome is 1 if the variable is relevant, 0 otherwise). Treating VI as if it were a 'prediction' for these binary outcomes yields a C-statistic which may be interpreted as the probability that the VI technique will rank a relevant variable higher than an irrelevant variable (Harrell et al., 1982).

4.2.6 Results

The three techniques that used 'aorsf' to estimate VI were ranked first (`aorsf-negate`; $C = 76.0$), second (`aorsf-anova`; $C = 74.0$), and third (`aorsf-permute`; $C = 73.3$) in overall mean C-statistic across all of the simulation scenarios, with `aorsf-negate` obtaining the highest C-statistic in 27 out of 36 VI tasks (Figure 6). Among the four relevant variable classes, `aorsf-negate` had the highest mean C-statistic for main effects, combination effects, and non-linear effects, with the greatest advantage of using `aorsf-negate` occurring among non-linear and combination variables. Full results from the experiment are provided in Table A.3. Computationally, ANOVA VI was faster than negation and permutation VI, with a median time of 1.34 seconds versus 12.3 and 14.1 seconds, respectively.

5. Discussion

In this paper, we have developed two contributions to the oblique RSF: (1) the accelerated oblique RSF (that is, `aorsf-fast`) and (2) negation VI. Our technique to accelerate the oblique RSF reduces the number of operations required to find linear combinations of inputs using a single iteration of Newton Raphson scoring, while our VI technique directly engages with coefficients in linear combinations of inputs to measure importance of individual variables. In numeric experiments, we found that that `aorsf-fast` is over 500 times faster and just as accurate in risk prediction tasks compared to its predecessor, `obliqueRSF-net`. We also found that negation VI, a technique to estimate VI using the oblique RSF, detected

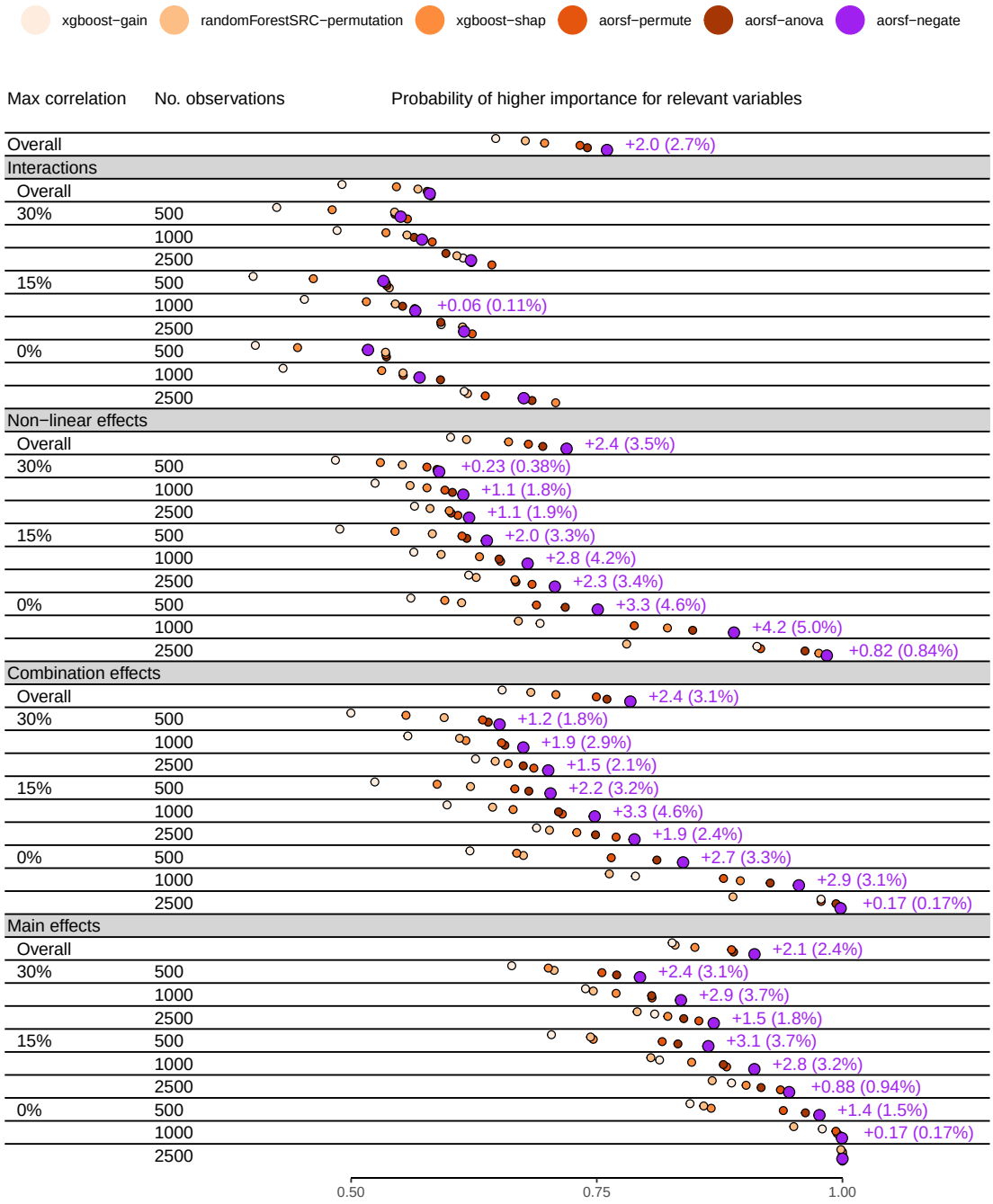

Figure 6: Concordance statistic for assigning higher importance to relevant versus irrelevant variables. Text appears in rows where negation importance obtained the highest concordance, showing absolute and percent improvement over the second best technique.

non-linear, combination, and main effects more effectively than three standard methods to estimate VI: permutation, ANOVA, and SHAP VI. Overall, we found that estimating VI using negation instead of ANOVA increased the C-statistic for ranking a relevant variable higher than an irrelevant variable by 2.00, a relative increase of 2.70%.

5.1 Implications of our results

Accurate risk prediction models have the potential to improve healthcare by directing timely interventions to patients who are most likely to benefit. However, prediction models that cannot scale adequately to large databases or cannot be interpreted and explained have no place in clinical practice. The current study advances the oblique RSF, an accurate risk prediction model, towards being accurate, scalable, and interpretable. The improved computational efficiency of the accelerated oblique RSF increases the feasibility of applying oblique RSFs in a wide range of prediction tasks. Faster model evaluation and re-fitting also improve diagnosis and resolution of model-based issues (for example, model calibration deteriorates over time). The introduction of negation VI also advances interpretability. VI is intrinsically linked to model fairness, as it can be used to identify when protected characteristics such as race, religion, and sexuality are inadvertently used (either directly or through correlates of these characteristics) by a prediction model. Since negation VI engages with the coefficients used in linear combinations of variables, a major component of oblique RSFs, it may be more capable of diagnosing unfairness in oblique RSFs compared to permutation importance and model-agnostic VI techniques.

5.2 Limitations and next steps

While the current study advances the oblique RSF towards being scalable and interpretable, there remain several limitations that can be targeted in future studies. The accelerated oblique RSF does not account for competing risks, and biased estimation of incidence may occur when competing risks are ignored. Thus, allowing the oblique RSF to account for competing risks is a high priority for future studies. In addition, missing data are not addressed in the accelerated oblique RSF, and users of the `aorsf` R package are expected to impute missing values prior to model training and testing. However, missing data are common and there are numerous techniques for ensemble tree methods to handle missing data during the tree growing procedure. Thus, a second item of high priority for future studies is to develop and evaluate strategies to handle missing data while growing an oblique RSF. Last, Cui et al. (2017) found that estimating an inverse-probability weighted hazard function at each non-leaf node of a survival tree allows the RSF to converge asymptotically to the true survival function when some variables contribute both to the risk of the event and the risk of censoring, a scenario that is very likely in the analysis of electronic medical records. The accelerated oblique RSF could incorporate this splitting technique by using Newton Raphson scoring to fit a model for the censoring distribution and then a weighted model could be fit to the failure distribution. This final item has the highest priority, as

Cui et al. (2017) showed it is a requisite condition for consistency of axis based survival trees in fairly general settings.

Acknowledgments

Research reported in this publication was supported by the Center for Biomedical Informatics, Wake Forest University School of Medicine. The project described was supported by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through Grant Award Number UL1TR001420. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Appendix

A.1: Data sets used for numeric experiments

Label	N observations	N predictors	Outcome	N Events	% Censored
VA lung cancer trial	137	8	Death	128	6.57
Colon cancer	929	12	Recurrence	468	49.6
			Death	452	51.3
Primary biliary cholangitis	276	19	Death	111	59.8
Movies released in 2015-2018	551	46	Gross 1M USD	522	5.26
GBSG II	686	10	Recurrence Or Death	299	56.4
Systolic Heart Failure	2,231	41	Death	726	67.5
Serum free light chain	7,874	10	Death	2,169	72.5
Non-alcohol fatty liver disease	17,549	24	Death	1,364	92.2
Rotterdam tumor bank	2,982	11	Recurrence	1,518	49.1
			Death	1,272	57.3
ACTG 320	1,151	12	AIDS Diagnosis	96	91.7
			Death	26	97.7
GUIDE-IT	894	59	Cardiovascular Death	110	87.7
			Hf Hospitalization	288	67.8
Early breast cancer	614	1,692	Recurrence Or Death	134	78.2
SPRINT	9,361	174	Cardiovascular Death	521	94.4
			Death	1,644	82.4
FCL	541	7	Death	76	86.0
			Relapse	272	49.7

Monoclonal gammopathy	1,384	8	Death	963	30.4
			Malignancy	115	91.7
MESA	6,783	48	Heart Failure	339	95.0
			Coronary Heart Disease	439	93.5
			Stroke	292	95.7
			Death	1,297	80.9
ARIC	13,623	41	Heart Failure	2,981	78.1
			Coronary Heart Disease	2,282	83.2
			Stroke	1,323	90.3
			Death	6,662	51.1

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks.

	Performance metric (SD)		Computation time, seconds	
	Scaled Brier	C-Statistic	Model fitting	Risk prediction
Overall				
aorsf-fast	0.136 (0.113)	0.776 (0.072)	0.721	0.174
aorsf-cph	0.136 (0.113)	0.775 (0.071)	1.867	0.169
aorsf-net	0.129 (0.128)	0.771 (0.075)	131.131	0.167
rsf-standard	0.123 (0.118)	0.761 (0.076)	2.116	0.278
cif-standard	0.123 (0.101)	0.769 (0.071)	4.312	14.590
obliqueRSF-net	0.122 (0.087)	0.775 (0.072)	379.917	23.183
cif-rotate	0.121 (0.130)	0.759 (0.082)	39.895	8.598
glmnet-cox	0.119 (0.124)	0.763 (0.075)	0.433	0.004
ranger-extratrees	0.107 (0.088)	0.765 (0.067)	1.986	1.289
cif-extension	0.102 (0.096)	0.764 (0.072)	22.524	7.926
aorsf-random	0.097 (0.084)	0.744 (0.064)	2.201	0.169
xgboost-cox	0.074 (0.104)	0.759 (0.089)	4.384	0.004
nn-cox	0.050 (0.109)	0.652 (0.138)	16.298	3.578
xgboost-aft	—	0.768 (0.075)	14.513	0.007
ACTG 320; AIDS diagnosis, n = 1151, p = 12				
ranger-extratrees	0.027 (0.017)	0.736 (0.036)	0.052	0.145
aorsf-random	0.027 (0.022)	0.741 (0.035)	0.408	0.035
obliqueRSF-net	0.026 (0.023)	0.741 (0.038)	26.825	15.427
cif-standard	0.023 (0.032)	0.740 (0.041)	1.062	4.542
aorsf-cph	0.023 (0.030)	0.744 (0.039)	0.436	0.035
cif-extension	0.023 (0.016)	0.719 (0.040)	9.253	4.180
aorsf-fast	0.022 (0.029)	0.737 (0.041)	0.142	0.034
aorsf-net	0.016 (0.035)	0.738 (0.040)	18.751	0.036
glmnet-cox	0.015 (0.029)	0.742 (0.035)	0.174	0.002
rsf-standard	0.005 (0.041)	0.727 (0.044)	0.149	0.059
cif-rotate	0.002 (0.042)	0.729 (0.040)	15.116	3.882
nn-cox	-0.001 (0.017)	0.636 (0.114)	10.863	0.537
xgboost-cox	-0.003 (0.046)	0.747 (0.031)	3.788	0.003
xgboost-aft	—	0.734 (0.034)	9.657	0.006
ACTG 320; death, n = 1151, p = 12				
obliqueRSF-net	0.007 (0.012)	0.824 (0.053)	8.814	10.613
aorsf-cph	0.006 (0.018)	0.823 (0.061)	0.351	0.020

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
aorsf-fast	0.006 (0.019)	0.829 (0.057)	0.086	0.020
aorsf-random	0.004 (0.015)	0.791 (0.075)	0.225	0.023
cif-extension	0.000 (0.020)	0.764 (0.067)	8.850	3.772
ranger-extratrees	0.000 (0.019)	0.776 (0.070)	0.041	0.124
xgboost-cox	-0.004 (0.004)	0.500 (0.000)	0.111	0.002
nn-cox	-0.005 (0.004)	0.521 (0.095)	10.159	0.520
cif-standard	-0.006 (0.025)	0.780 (0.063)	1.117	4.526
aorsf-net	-0.006 (0.033)	0.808 (0.068)	14.237	0.023
rsf-standard	-0.033 (0.052)	0.776 (0.075)	0.094	0.036
cif-rotate	-0.040 (0.048)	0.702 (0.090)	13.345	3.292
glmnet-cox	-0.065 (0.098)	0.747 (0.100)	0.276	0.002
xgboost-aft	—	0.773 (0.072)	9.772	0.006
ARIC; coronary heart disease, n = 13623, p = 41				
aorsf-fast	0.157 (0.007)	0.810 (0.007)	4.561	1.355
aorsf-cph	0.153 (0.007)	0.809 (0.007)	14.677	1.379
aorsf-net	0.153 (0.007)	0.810 (0.007)	511.536	1.484
rsf-standard	0.151 (0.007)	0.801 (0.007)	9.099	1.077
obliqueRSF-net	0.144 (0.005)	0.812 (0.007)	2814.937	353.143
cif-standard	0.132 (0.005)	0.810 (0.007)	70.476	358.500
glmnet-cox	0.129 (0.012)	0.796 (0.008)	1.547	0.011
nn-cox	0.122 (0.014)	0.795 (0.007)	52.898	105.741
ranger-extratrees	0.112 (0.005)	0.796 (0.009)	305.330	66.584
cif-rotate	0.104 (0.004)	0.784 (0.007)	571.763	70.690
aorsf-random	0.098 (0.005)	0.772 (0.008)	11.675	1.360
cif-extension	0.069 (0.002)	0.787 (0.009)	165.124	51.252
xgboost-cox	0.065 (0.017)	0.814 (0.005)	8.516	0.015
xgboost-aft	—	0.815 (0.006)	22.582	0.013
ARIC; death, n = 13623, p = 41				
aorsf-net	0.217 (0.006)	0.792 (0.004)	955.499	2.412
rsf-standard	0.216 (0.007)	0.789 (0.004)	13.881	1.320
aorsf-cph	0.215 (0.007)	0.792 (0.004)	23.030	2.291
aorsf-fast	0.215 (0.007)	0.792 (0.004)	7.760	2.281
obliqueRSF-net	0.207 (0.005)	0.792 (0.004)	7413.571	319.191
cif-standard	0.202 (0.004)	0.790 (0.004)	71.453	386.302
nn-cox	0.192 (0.012)	0.780 (0.005)	97.331	99.624

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
glmnet-cox	0.190 (0.015)	0.777 (0.008)	2.123	0.012
ranger-extratrees	0.181 (0.004)	0.781 (0.005)	406.320	66.433
cif-rotate	0.152 (0.007)	0.757 (0.006)	581.210	65.820
xgboost-cox	0.132 (0.011)	0.795 (0.004)	12.310	0.016
aorsf-random	0.130 (0.005)	0.733 (0.005)	21.733	2.051
cif-extension	0.113 (0.002)	0.776 (0.005)	183.959	53.651
xgboost-aft	—	0.794 (0.004)	27.733	0.014
ARIC; heart failure, n = 13623, p = 41				
aorsf-fast	0.234 (0.007)	0.842 (0.005)	5.445	1.619
rsf-standard	0.230 (0.007)	0.835 (0.005)	10.089	1.049
aorsf-cph	0.229 (0.007)	0.841 (0.005)	16.936	1.624
aorsf-net	0.217 (0.057)	0.832 (0.043)	631.093	1.707
obliqueRSF-net	0.213 (0.005)	0.841 (0.005)	3701.663	311.291
cif-standard	0.199 (0.005)	0.839 (0.005)	74.710	369.446
nn-cox	0.187 (0.020)	0.826 (0.009)	59.242	95.742
cif-rotate	0.172 (0.006)	0.806 (0.007)	581.038	67.850
ranger-extratrees	0.171 (0.004)	0.824 (0.005)	326.124	86.921
glmnet-cox	0.164 (0.046)	0.816 (0.019)	2.424	0.012
aorsf-random	0.139 (0.005)	0.789 (0.006)	15.571	1.574
xgboost-cox	0.121 (0.018)	0.845 (0.005)	11.223	0.015
cif-extension	0.109 (0.003)	0.809 (0.006)	170.011	49.372
xgboost-aft	—	0.844 (0.005)	26.160	0.014
ARIC; stroke, n = 13623, p = 41				
aorsf-fast	0.093 (0.004)	0.794 (0.007)	4.101	1.127
rsf-standard	0.090 (0.006)	0.785 (0.007)	7.258	0.919
aorsf-cph	0.090 (0.004)	0.793 (0.007)	13.277	1.119
aorsf-net	0.087 (0.017)	0.786 (0.031)	388.977	1.200
obliqueRSF-net	0.082 (0.003)	0.792 (0.007)	1768.748	384.697
glmnet-cox	0.078 (0.005)	0.788 (0.007)	1.662	0.010
cif-standard	0.073 (0.003)	0.788 (0.007)	72.160	363.855
nn-cox	0.071 (0.013)	0.781 (0.009)	31.997	88.257
ranger-extratrees	0.067 (0.003)	0.780 (0.008)	249.618	67.732
aorsf-random	0.059 (0.004)	0.751 (0.009)	9.374	1.121
cif-rotate	0.052 (0.003)	0.769 (0.009)	592.458	70.175
xgboost-cox	0.047 (0.013)	0.794 (0.006)	6.688	0.014

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
cif-extension	0.036 (0.002)	0.771 (0.008)	170.852	51.402
xgboost-aft	—	0.793 (0.007)	21.053	0.013
Colon cancer; death, n = 929, p = 12				
aorsf-fast	0.098 (0.013)	0.718 (0.012)	0.228	0.050
aorsf-cph	0.097 (0.013)	0.717 (0.012)	0.630	0.051
cif-standard	0.096 (0.013)	0.710 (0.012)	0.670	3.620
aorsf-random	0.094 (0.009)	0.715 (0.012)	0.992	0.045
aorsf-net	0.090 (0.024)	0.710 (0.036)	49.564	0.046
obliqueRSF-net	0.088 (0.006)	0.717 (0.012)	250.492	16.952
cif-rotate	0.086 (0.018)	0.705 (0.015)	12.341	3.533
rsf-standard	0.085 (0.019)	0.703 (0.011)	1.557	0.151
ranger-extratrees	0.082 (0.007)	0.710 (0.012)	0.089	0.241
cif-extension	0.080 (0.006)	0.709 (0.012)	8.368	3.934
glmnet-cox	0.072 (0.013)	0.709 (0.018)	0.104	0.003
xgboost-cox	0.061 (0.012)	0.700 (0.013)	3.143	0.003
nn-cox	-0.003 (0.003)	0.515 (0.030)	11.966	1.368
xgboost-aft	—	0.706 (0.014)	11.263	0.006
Colon cancer; recurrence, n = 929, p = 12				
aorsf-fast	0.099 (0.017)	0.713 (0.016)	0.229	0.050
aorsf-cph	0.098 (0.016)	0.712 (0.015)	0.639	0.051
cif-standard	0.090 (0.016)	0.701 (0.017)	0.665	3.411
aorsf-net	0.087 (0.042)	0.708 (0.026)	50.808	0.048
obliqueRSF-net	0.087 (0.009)	0.710 (0.016)	258.163	16.507
aorsf-random	0.086 (0.013)	0.702 (0.016)	1.023	0.047
cif-rotate	0.084 (0.021)	0.695 (0.018)	12.248	3.554
cif-extension	0.081 (0.009)	0.706 (0.017)	8.129	3.936
rsf-standard	0.080 (0.021)	0.694 (0.015)	1.654	0.147
ranger-extratrees	0.079 (0.011)	0.699 (0.016)	0.090	0.242
glmnet-cox	0.071 (0.018)	0.704 (0.024)	0.116	0.003
xgboost-cox	0.059 (0.011)	0.694 (0.018)	3.002	0.003
nn-cox	-0.014 (0.032)	0.503 (0.034)	13.121	1.248
xgboost-aft	—	0.701 (0.019)	12.035	0.006
Early breast cancer; recurrence or death, n = 614, p = 1692				
obliqueRSF-net	0.073 (0.023)	0.750 (0.027)	1867.035	13.505

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
cif-rotate	0.071 (0.017)	0.746 (0.027)	6498.327	347.595
cif-standard	0.068 (0.017)	0.746 (0.028)	8.392	4.008
aorsf-cph	0.066 (0.030)	0.744 (0.022)	1.313	0.182
aorsf-fast	0.065 (0.029)	0.745 (0.023)	0.740	0.180
cif-extension	0.065 (0.014)	0.745 (0.025)	43.993	6.218
ranger-extratrees	0.061 (0.022)	0.741 (0.028)	0.218	0.676
glmnet-cox	0.044 (0.032)	0.724 (0.034)	5.824	0.006
aorsf-random	0.027 (0.014)	0.694 (0.038)	1.692	0.172
xgboost-cox	0.027 (0.034)	0.740 (0.027)	2.297	0.007
rsf-standard	0.021 (0.039)	0.692 (0.032)	0.777	0.314
nn-cox	-0.017 (0.073)	0.681 (0.048)	16.971	1.662
aorsf-net	-0.039 (0.249)	0.731 (0.052)	455.331	0.176
xgboost-aft	—	0.741 (0.023)	9.590	0.010
FCL; death, n = 541, p = 7				
glmnet-cox	0.116 (0.029)	0.787 (0.037)	0.091	0.002
aorsf-fast	0.100 (0.040)	0.768 (0.033)	0.079	0.019
aorsf-cph	0.100 (0.041)	0.769 (0.033)	0.165	0.019
aorsf-net	0.093 (0.043)	0.752 (0.049)	13.137	0.019
obliqueRSF-net	0.090 (0.028)	0.761 (0.037)	96.051	5.285
cif-extension	0.086 (0.038)	0.731 (0.036)	5.213	2.395
aorsf-random	0.086 (0.031)	0.757 (0.033)	0.259	0.019
cif-rotate	0.085 (0.050)	0.756 (0.028)	6.045	2.042
cif-standard	0.085 (0.040)	0.747 (0.035)	0.289	1.106
ranger-extratrees	0.073 (0.017)	0.742 (0.032)	0.043	0.083
rsf-standard	0.071 (0.051)	0.733 (0.036)	0.111	0.038
xgboost-cox	0.033 (0.052)	0.697 (0.109)	0.374	0.002
nn-cox	-0.003 (0.022)	0.534 (0.114)	11.145	0.437
xgboost-aft	—	0.756 (0.039)	7.831	0.006
FCL; relapse, n = 541, p = 7				
glmnet-cox	0.027 (0.018)	0.619 (0.025)	0.089	0.002
ranger-extratrees	0.016 (0.015)	0.594 (0.025)	0.032	0.080
obliqueRSF-net	0.011 (0.016)	0.589 (0.023)	217.918	6.187
aorsf-random	0.011 (0.018)	0.594 (0.025)	0.395	0.021
xgboost-cox	0.010 (0.016)	0.596 (0.031)	1.291	0.002
aorsf-fast	0.005 (0.020)	0.592 (0.026)	0.111	0.022

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
aorsf-cph	0.005 (0.021)	0.593 (0.027)	0.262	0.022
cif-standard	0.005 (0.020)	0.591 (0.022)	0.290	1.120
aorsf-net	0.005 (0.021)	0.590 (0.027)	19.255	0.022
nn-cox	-0.003 (0.023)	0.544 (0.064)	11.514	0.442
cif-extension	-0.007 (0.023)	0.576 (0.028)	6.044	2.303
cif-rotate	-0.015 (0.025)	0.580 (0.030)	6.986	1.965
rsf-standard	-0.030 (0.032)	0.575 (0.025)	0.843	0.086
xgboost-aft	—	0.580 (0.035)	6.226	0.006
GBSG II; recurrence or death, n = 686, p = 10				
obliqueRSF-net	0.124 (0.016)	0.746 (0.018)	307.849	6.881
cif-standard	0.123 (0.020)	0.743 (0.020)	0.430	1.769
rsf-standard	0.120 (0.024)	0.737 (0.020)	1.508	0.116
aorsf-cph	0.120 (0.026)	0.736 (0.018)	0.406	0.038
aorsf-fast	0.117 (0.025)	0.733 (0.017)	0.168	0.039
cif-extension	0.115 (0.018)	0.743 (0.020)	7.990	3.331
aorsf-net	0.112 (0.056)	0.734 (0.036)	37.488	0.038
cif-rotate	0.107 (0.024)	0.729 (0.018)	10.801	2.745
aorsf-random	0.105 (0.025)	0.724 (0.026)	0.785	0.036
ranger-extratrees	0.095 (0.018)	0.738 (0.025)	0.052	0.136
glmnet-cox	0.090 (0.018)	0.728 (0.021)	0.098	0.002
xgboost-cox	0.082 (0.017)	0.730 (0.019)	2.465	0.003
nn-cox	-0.003 (0.005)	0.521 (0.051)	11.570	0.838
xgboost-aft	—	0.730 (0.022)	10.495	0.006
GUIDE-IT; CVD death, n = 894, p = 59				
aorsf-fast	0.076 (0.017)	0.748 (0.028)	0.160	0.037
aorsf-net	0.075 (0.017)	0.746 (0.028)	27.437	0.039
aorsf-cph	0.072 (0.018)	0.744 (0.029)	0.379	0.037
glmnet-cox	0.068 (0.040)	0.721 (0.084)	0.510	0.003
obliqueRSF-net	0.063 (0.013)	0.742 (0.024)	222.174	11.114
cif-rotate	0.060 (0.013)	0.721 (0.027)	35.135	5.270
cif-standard	0.059 (0.013)	0.739 (0.023)	1.108	3.576
ranger-extratrees	0.054 (0.012)	0.740 (0.029)	0.091	0.183
cif-extension	0.052 (0.011)	0.730 (0.023)	13.427	5.518
rsf-standard	0.048 (0.021)	0.707 (0.026)	0.169	0.060
xgboost-cox	0.045 (0.045)	0.746 (0.019)	3.514	0.003

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
aorsf-random	0.033 (0.013)	0.700 (0.029)	0.511	0.039
nn-cox	0.009 (0.020)	0.628 (0.086)	10.335	0.565
xgboost-aft	—	0.733 (0.021)	10.963	0.006
GUIDE-IT; HF hospitalization, n = 894, p = 59				
aorsf-cph	0.080 (0.018)	0.721 (0.024)	0.683	0.054
aorsf-fast	0.079 (0.019)	0.721 (0.026)	0.241	0.053
obliqueRSF-net	0.072 (0.011)	0.720 (0.024)	385.158	9.315
ranger-extratrees	0.072 (0.010)	0.721 (0.023)	0.236	0.189
aorsf-net	0.072 (0.041)	0.715 (0.036)	53.536	0.057
cif-standard	0.069 (0.010)	0.715 (0.024)	0.931	3.280
cif-rotate	0.066 (0.019)	0.706 (0.030)	41.946	5.179
cif-extension	0.063 (0.009)	0.713 (0.023)	14.574	5.603
glmnet-cox	0.058 (0.020)	0.699 (0.025)	0.462	0.003
rsf-standard	0.055 (0.022)	0.692 (0.026)	1.528	0.119
aorsf-random	0.048 (0.010)	0.682 (0.023)	0.920	0.053
nn-cox	0.046 (0.027)	0.695 (0.040)	12.586	0.590
xgboost-cox	0.038 (0.017)	0.697 (0.027)	2.920	0.003
xgboost-aft	—	0.695 (0.026)	13.042	0.006
MESA; coronary heart disease, n = 6785, p = 48				
aorsf-fast	0.064 (0.009)	0.807 (0.010)	1.254	0.359
aorsf-net	0.063 (0.010)	0.804 (0.010)	175.507	0.380
obliqueRSF-net	0.063 (0.007)	0.808 (0.009)	495.036	263.736
aorsf-cph	0.061 (0.009)	0.801 (0.011)	5.065	0.365
cif-standard	0.059 (0.006)	0.803 (0.010)	23.165	98.928
cif-rotate	0.059 (0.007)	0.802 (0.010)	291.147	38.590
rsf-standard	0.058 (0.011)	0.794 (0.011)	3.497	1.115
ranger-extratrees	0.047 (0.004)	0.794 (0.010)	8.046	7.095
cif-extension	0.047 (0.003)	0.805 (0.010)	100.181	30.035
glmnet-cox	0.039 (0.016)	0.775 (0.015)	4.983	0.007
nn-cox	0.034 (0.014)	0.771 (0.018)	17.346	16.923
aorsf-random	0.031 (0.005)	0.734 (0.015)	3.611	0.409
xgboost-cox	0.016 (0.022)	0.804 (0.010)	4.416	0.008
xgboost-aft	—	0.803 (0.010)	18.681	0.009
MESA; death, n = 6793, p = 48				

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
aorsf-net	0.145 (0.008)	0.792 (0.009)	304.543	0.545
aorsf-fast	0.145 (0.009)	0.792 (0.009)	1.769	0.625
aorsf-cph	0.143 (0.008)	0.791 (0.009)	6.747	0.522
rsf-standard	0.141 (0.008)	0.784 (0.009)	4.694	0.475
obliqueRSF-net	0.140 (0.007)	0.791 (0.009)	1154.938	154.248
nn-cox	0.134 (0.012)	0.790 (0.010)	31.798	18.891
cif-standard	0.134 (0.007)	0.788 (0.009)	23.348	100.276
glmnet-cox	0.129 (0.027)	0.789 (0.012)	1.504	0.007
cif-rotate	0.126 (0.007)	0.784 (0.010)	319.140	36.836
ranger-extratrees	0.114 (0.004)	0.784 (0.008)	7.843	6.084
cif-extension	0.092 (0.003)	0.781 (0.009)	110.957	30.312
aorsf-random	0.069 (0.004)	0.725 (0.008)	5.765	0.551
xgboost-cox	0.056 (0.029)	0.794 (0.009)	8.347	0.009
xgboost-aft	—	0.793 (0.009)	21.600	0.009
MESA; heart failure, n = 6785, p = 48				
aorsf-fast	0.114 (0.010)	0.867 (0.012)	1.145	0.317
aorsf-cph	0.109 (0.011)	0.860 (0.014)	4.847	0.330
rsf-standard	0.108 (0.011)	0.857 (0.010)	3.069	1.095
obliqueRSF-net	0.106 (0.008)	0.870 (0.012)	403.026	336.427
cif-rotate	0.104 (0.010)	0.870 (0.012)	262.282	37.519
cif-standard	0.102 (0.009)	0.865 (0.012)	24.266	98.770
aorsf-net	0.099 (0.067)	0.851 (0.065)	152.494	0.342
cif-extension	0.076 (0.005)	0.865 (0.011)	94.286	30.425
ranger-extratrees	0.075 (0.005)	0.849 (0.015)	7.052	6.710
nn-cox	0.071 (0.026)	0.828 (0.022)	15.582	14.999
aorsf-random	0.063 (0.006)	0.795 (0.015)	2.638	0.368
glmnet-cox	0.044 (0.044)	0.761 (0.146)	3.775	0.007
xgboost-cox	-0.010 (0.020)	0.871 (0.009)	6.431	0.009
xgboost-aft	—	0.872 (0.012)	20.098	0.009
MESA; stroke, n = 6783, p = 48				
cif-rotate	0.025 (0.005)	0.762 (0.017)	271.158	37.410
obliqueRSF-net	0.025 (0.004)	0.764 (0.016)	354.662	286.074
cif-standard	0.025 (0.004)	0.760 (0.017)	23.751	98.574
aorsf-fast	0.024 (0.006)	0.762 (0.015)	1.095	0.311
aorsf-cph	0.023 (0.005)	0.756 (0.016)	4.436	0.320

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
ranger-extratrees	0.022 (0.003)	0.757 (0.015)	6.801	6.789
glmnet-cox	0.021 (0.010)	0.764 (0.018)	3.726	0.007
cif-extension	0.021 (0.002)	0.767 (0.018)	95.584	29.197
rsf-standard	0.019 (0.009)	0.744 (0.018)	3.147	1.066
aorsf-net	0.018 (0.026)	0.749 (0.043)	138.878	0.330
nn-cox	0.014 (0.010)	0.725 (0.051)	17.348	20.132
aorsf-random	0.013 (0.003)	0.711 (0.023)	2.519	0.338
xgboost-cox	-0.001 (0.025)	0.761 (0.018)	4.044	0.008
xgboost-aft	—	0.763 (0.016)	17.880	0.010
Monoclonal gammopathy; death, n = 1384, p = 8				
cif-rotate	0.159 (0.019)	0.744 (0.015)	15.254	4.543
aorsf-cph	0.157 (0.016)	0.742 (0.011)	1.155	0.087
aorsf-fast	0.156 (0.016)	0.742 (0.011)	0.393	0.088
obliqueRSF-net	0.155 (0.013)	0.743 (0.011)	229.681	12.810
aorsf-net	0.154 (0.015)	0.741 (0.011)	88.452	0.086
cif-standard	0.150 (0.015)	0.738 (0.012)	0.989	5.746
rsf-standard	0.150 (0.017)	0.736 (0.011)	2.022	0.201
aorsf-random	0.145 (0.014)	0.734 (0.012)	1.762	0.082
cif-extension	0.143 (0.010)	0.747 (0.013)	10.829	4.681
glmnet-cox	0.139 (0.021)	0.728 (0.014)	0.118	0.003
xgboost-cox	0.123 (0.012)	0.733 (0.012)	3.847	0.003
ranger-extratrees	0.115 (0.005)	0.744 (0.012)	0.060	0.184
nn-cox	0.028 (0.043)	0.606 (0.100)	15.492	0.691
xgboost-aft	—	0.733 (0.013)	12.513	0.006
Monoclonal gammopathy; malignancy, n = 1384, p = 8				
glmnet-cox	0.015 (0.012)	0.648 (0.058)	0.104	0.002
aorsf-cph	0.010 (0.011)	0.644 (0.033)	0.592	0.041
aorsf-fast	0.010 (0.012)	0.640 (0.033)	0.192	0.042
ranger-extratrees	0.008 (0.006)	0.642 (0.031)	0.050	0.175
cif-extension	0.008 (0.008)	0.625 (0.025)	8.953	4.503
aorsf-net	0.007 (0.012)	0.641 (0.031)	22.912	0.042
aorsf-random	0.007 (0.013)	0.634 (0.032)	0.510	0.041
cif-standard	0.006 (0.009)	0.625 (0.028)	1.137	6.173
xgboost-cox	0.005 (0.017)	0.638 (0.038)	1.774	0.003
obliqueRSF-net	0.003 (0.022)	0.624 (0.038)	41.561	16.686

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
nn-cox	-0.003 (0.006)	0.507 (0.055)	10.592	0.704
rsf-standard	-0.010 (0.015)	0.614 (0.029)	0.744	0.074
cif-rotate	-0.024 (0.022)	0.550 (0.032)	12.665	4.579
xgboost-aft	—	0.627 (0.039)	11.031	0.006
Movies released in 2015-2018; gross 1M USD, n = 551, p = 46				
cif-rotate	0.635 (0.025)	0.943 (0.007)	19.464	3.497
glmnet-cox	0.616 (0.036)	0.940 (0.009)	0.179	0.002
aorsf-net	0.529 (0.029)	0.928 (0.011)	51.788	0.044
aorsf-cph	0.520 (0.024)	0.925 (0.011)	0.783	0.042
rsf-standard	0.518 (0.022)	0.922 (0.011)	1.324	0.103
aorsf-fast	0.514 (0.028)	0.922 (0.013)	0.216	0.043
xgboost-cox	0.514 (0.029)	0.932 (0.009)	13.877	0.004
nn-cox	0.512 (0.068)	0.901 (0.027)	17.133	0.681
cif-standard	0.469 (0.028)	0.901 (0.018)	0.375	1.238
cif-extension	0.452 (0.026)	0.919 (0.013)	9.070	3.976
ranger-extratrees	0.428 (0.026)	0.898 (0.019)	0.047	0.108
obliqueRSF-net	0.318 (0.023)	0.908 (0.018)	156.326	9.034
aorsf-random	0.299 (0.033)	0.849 (0.029)	0.901	0.039
xgboost-aft	—	0.927 (0.010)	35.252	0.007
Non-alcohol fatty liver disease; death, n = 17549, p = 24				
aorsf-cph	0.214 (0.008)	0.869 (0.005)	17.839	1.266
aorsf-fast	0.213 (0.008)	0.870 (0.005)	4.838	1.314
aorsf-net	0.211 (0.007)	0.865 (0.006)	453.851	1.297
obliqueRSF-net	0.210 (0.008)	0.869 (0.005)	1414.334	1054.087
rsf-standard	0.208 (0.009)	0.860 (0.005)	9.897	1.177
glmnet-cox	0.207 (0.010)	0.861 (0.005)	1.745	0.012
cif-standard	0.206 (0.007)	0.864 (0.006)	67.606	629.975
cif-rotate	0.191 (0.007)	0.866 (0.005)	262.207	61.085
ranger-extratrees	0.181 (0.007)	0.861 (0.005)	37.518	96.239
cif-extension	0.167 (0.003)	0.867 (0.006)	125.152	53.538
aorsf-random	0.141 (0.007)	0.840 (0.006)	9.678	1.364
xgboost-cox	0.021 (0.015)	0.877 (0.005)	9.002	0.018
nn-cox	0.000 (0.000)	0.555 (0.092)	21.841	115.006
xgboost-aft	—	0.875 (0.005)	29.487	0.014

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
Primary biliary cholangitis; death, n = 276, p = 19				
aorsf-fast	0.431 (0.033)	0.907 (0.022)	0.069	0.018
aorsf-cph	0.419 (0.034)	0.905 (0.022)	0.152	0.018
cif-rotate	0.403 (0.042)	0.897 (0.022)	9.844	2.217
aorsf-net	0.396 (0.078)	0.899 (0.031)	14.684	0.019
rsf-standard	0.393 (0.036)	0.894 (0.024)	0.094	0.038
obliqueRSF-net	0.369 (0.034)	0.905 (0.023)	110.615	1.764
aorsf-random	0.356 (0.029)	0.894 (0.021)	0.286	0.019
cif-standard	0.351 (0.035)	0.902 (0.026)	0.186	0.363
cif-extension	0.348 (0.035)	0.899 (0.024)	5.248	2.252
glmnet-cox	0.340 (0.047)	0.885 (0.029)	0.115	0.002
ranger-extratrees	0.275 (0.028)	0.893 (0.027)	0.028	0.037
xgboost-cox	0.249 (0.106)	0.878 (0.026)	4.873	0.003
nn-cox	-0.017 (0.017)	0.551 (0.130)	10.566	0.240
xgboost-aft	—	0.882 (0.024)	10.124	0.006
Rotterdam tumor bank; death, n = 2982, p = 11				
aorsf-net	0.165 (0.012)	0.761 (0.009)	156.780	0.189
obliqueRSF-net	0.162 (0.010)	0.760 (0.009)	449.043	37.754
aorsf-cph	0.162 (0.012)	0.758 (0.009)	2.515	0.198
aorsf-fast	0.159 (0.013)	0.756 (0.009)	0.796	0.198
cif-standard	0.159 (0.011)	0.758 (0.009)	4.650	22.504
rsf-standard	0.157 (0.014)	0.755 (0.009)	2.710	0.802
aorsf-random	0.153 (0.011)	0.751 (0.010)	2.990	0.184
cif-rotate	0.148 (0.012)	0.751 (0.012)	34.249	8.476
ranger-extratrees	0.138 (0.006)	0.748 (0.009)	2.921	2.599
xgboost-cox	0.131 (0.013)	0.753 (0.010)	3.742	0.004
cif-extension	0.130 (0.004)	0.751 (0.009)	22.114	8.625
glmnet-cox	0.118 (0.008)	0.732 (0.009)	0.212	0.004
nn-cox	-0.032 (0.092)	0.513 (0.042)	16.455	8.672
xgboost-aft	—	0.760 (0.009)	14.327	0.007
Rotterdam tumor bank; recurrence, n = 2982, p = 11				
obliqueRSF-net	0.148 (0.011)	0.737 (0.010)	529.355	39.185
aorsf-net	0.146 (0.012)	0.735 (0.009)	166.455	0.193
cif-standard	0.144 (0.011)	0.734 (0.009)	4.545	22.929

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
aorsf-cph	0.144 (0.012)	0.734 (0.009)	2.760	0.211
aorsf-fast	0.143 (0.012)	0.733 (0.009)	0.872	0.210
aorsf-random	0.140 (0.011)	0.730 (0.009)	3.281	0.190
rsf-standard	0.138 (0.012)	0.731 (0.009)	2.861	0.846
ranger-extratrees	0.135 (0.007)	0.734 (0.009)	2.765	2.677
cif-rotate	0.129 (0.010)	0.725 (0.009)	36.702	8.632
cif-extension	0.119 (0.006)	0.731 (0.009)	22.734	8.788
glmnet-cox	0.117 (0.008)	0.727 (0.009)	0.227	0.004
xgboost-cox	0.114 (0.008)	0.729 (0.009)	3.255	0.004
nn-cox	-0.011 (0.030)	0.500 (0.055)	17.957	9.371
xgboost-aft	—	0.735 (0.009)	14.774	0.007
Serum free light chain; death, n = 7874, p = 10				
aorsf-fast	0.250 (0.014)	0.826 (0.008)	2.028	0.592
aorsf-cph	0.250 (0.014)	0.825 (0.008)	6.526	0.593
glmnet-cox	0.247 (0.012)	0.820 (0.007)	0.486	0.006
obliqueRSF-net	0.247 (0.012)	0.821 (0.008)	1094.380	148.463
aorsf-net	0.246 (0.024)	0.821 (0.014)	293.468	0.566
ranger-extratrees	0.243 (0.009)	0.820 (0.007)	9.981	11.444
rsf-standard	0.243 (0.014)	0.815 (0.008)	4.619	0.561
cif-standard	0.243 (0.011)	0.818 (0.008)	19.411	119.478
aorsf-random	0.232 (0.012)	0.817 (0.008)	6.725	0.570
cif-rotate	0.227 (0.008)	0.819 (0.007)	65.281	21.263
cif-extension	0.201 (0.005)	0.820 (0.008)	39.592	20.428
xgboost-cox	0.094 (0.040)	0.823 (0.008)	5.798	0.008
nn-cox	0.001 (0.006)	0.591 (0.114)	23.551	26.323
xgboost-aft	—	0.823 (0.008)	18.228	0.009
SPRINT; CVD death, n = 9361, p = 174				
glmnet-cox	0.070 (0.011)	0.794 (0.010)	13.586	0.010
aorsf-net	0.070 (0.007)	0.796 (0.011)	348.986	0.675
aorsf-fast	0.069 (0.006)	0.796 (0.011)	2.415	0.604
aorsf-cph	0.069 (0.006)	0.796 (0.011)	9.014	0.612
obliqueRSF-net	0.067 (0.005)	0.797 (0.012)	1014.547	425.668
rsf-standard	0.064 (0.007)	0.787 (0.014)	4.574	1.245
cif-standard	0.061 (0.004)	0.797 (0.011)	51.137	182.954
cif-rotate	0.061 (0.005)	0.789 (0.012)	971.559	115.147

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
ranger-extratrees	0.054 (0.003)	0.791 (0.012)	7.589	8.207
nn-cox	0.036 (0.020)	0.767 (0.019)	21.632	28.604
cif-extension	0.034 (0.002)	0.788 (0.011)	124.094	32.778
aorsf-random	0.026 (0.003)	0.746 (0.017)	5.745	0.743
xgboost-cox	0.007 (0.018)	0.798 (0.011)	6.861	0.014
xgboost-aft	—	0.794 (0.012)	21.141	0.013
SPRINT; death, n = 9361, p = 174				
glmnet-cox	0.122 (0.011)	0.770 (0.010)	5.120	0.011
aorsf-cph	0.116 (0.008)	0.770 (0.009)	13.398	1.453
aorsf-fast	0.116 (0.008)	0.770 (0.008)	3.686	1.431
aorsf-net	0.113 (0.009)	0.768 (0.009)	609.415	0.984
obliqueRSF-net	0.112 (0.007)	0.767 (0.008)	2609.859	237.782
rsf-standard	0.110 (0.007)	0.762 (0.009)	6.441	0.687
cif-standard	0.106 (0.006)	0.764 (0.008)	49.213	190.044
nn-cox	0.098 (0.010)	0.757 (0.010)	36.524	31.856
ranger-extratrees	0.096 (0.005)	0.756 (0.009)	10.926	9.128
cif-rotate	0.090 (0.006)	0.745 (0.009)	1072.994	115.077
cif-extension	0.055 (0.002)	0.746 (0.009)	136.086	33.237
aorsf-random	0.052 (0.003)	0.719 (0.009)	9.318	1.018
xgboost-cox	0.030 (0.024)	0.772 (0.008)	9.349	0.013
xgboost-aft	—	0.771 (0.008)	26.417	0.014
Systolic Heart Failure; death, n = 2231, p = 41				
obliqueRSF-net	0.114 (0.012)	0.748 (0.012)	379.908	24.724
glmnet-cox	0.114 (0.013)	0.747 (0.012)	0.260	0.003
cif-rotate	0.114 (0.013)	0.742 (0.011)	70.063	10.518
aorsf-net	0.113 (0.013)	0.745 (0.012)	120.016	0.160
aorsf-cph	0.112 (0.014)	0.746 (0.012)	1.972	0.152
cif-standard	0.110 (0.011)	0.745 (0.012)	3.971	15.781
aorsf-fast	0.110 (0.016)	0.745 (0.011)	0.617	0.153
rsf-standard	0.106 (0.011)	0.737 (0.010)	2.067	0.278
cif-extension	0.095 (0.005)	0.745 (0.012)	28.919	9.534
ranger-extratrees	0.092 (0.008)	0.739 (0.013)	2.936	1.775
xgboost-cox	0.091 (0.009)	0.746 (0.010)	4.368	0.004
aorsf-random	0.082 (0.005)	0.733 (0.012)	2.553	0.149
nn-cox	0.078 (0.024)	0.710 (0.031)	18.849	4.563

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
xgboost-aft	—	0.743 (0.009)	13.559	0.007
VA lung cancer trial; death, n = 137, p = 8				
aorsf-fast	0.197 (0.050)	0.794 (0.033)	0.047	0.011
aorsf-net	0.197 (0.050)	0.796 (0.034)	9.985	0.012
aorsf-cph	0.194 (0.053)	0.793 (0.035)	0.097	0.012
cif-rotate	0.192 (0.066)	0.786 (0.037)	4.248	1.059
rsf-standard	0.172 (0.049)	0.786 (0.038)	0.067	0.025
cif-extension	0.169 (0.048)	0.793 (0.032)	3.508	1.194
glmnet-cox	0.157 (0.032)	0.786 (0.038)	0.078	0.002
aorsf-random	0.151 (0.046)	0.779 (0.037)	0.205	0.011
cif-standard	0.124 (0.040)	0.767 (0.035)	0.098	0.119
obliqueRSF-net	0.123 (0.034)	0.794 (0.028)	59.944	0.669
ranger-extratrees	0.089 (0.033)	0.777 (0.036)	0.021	0.026
xgboost-cox	0.057 (0.076)	0.748 (0.048)	1.099	0.002
xgboost-aft	—	0.755 (0.045)	5.540	0.005
nn-cox	-0.031 (0.036)	0.523 (0.080)	11.159	0.126

A.3: Discrimination of relevant versus irrelevant variables for several techniques to estimate variable importance.

		accelerated oblique RSF			xgboost		RSF
Max correlation	No. observations	Negation	ANOVA	Permutation	SHAP	Gain	Permutation
Overall	Overall	76.0	74.0	73.3	69.7	64.7	67.7
Interactions							
Overall	Overall	58.0	57.7	58.1	54.6	49.1	56.8
30	500	55.0	54.5	55.7	48.1	42.4	54.4
30	1,000	57.2	56.4	58.3	53.5	48.6	55.7
30	2,500	62.2	59.6	64.3	62.2	61.4	60.8
15	500	53.3	53.6	53.6	46.2	40.0	53.9
15	1,000	56.5	55.2	56.5	51.6	45.2	54.5
15	2,500	61.5	59.1	62.3	61.5	59.2	61.3
0	500	51.7	53.6	53.5	44.6	40.3	53.5
0	1,000	57.0	59.1	55.3	53.1	43.1	55.3
0	2,500	67.6	68.4	63.7	70.8	61.5	61.8
Non-linear effects							
Overall	Overall	71.9	69.5	68.0	66.0	60.1	61.7
30	500	59.0	58.7	57.7	53.0	48.4	55.2
30	1,000	61.4	60.3	59.5	57.7	52.4	56.0
30	2,500	62.0	60.2	60.9	60.0	56.5	58.0
15	500	63.8	61.8	61.3	54.5	48.9	58.3
15	1,000	68.0	65.1	65.2	63.1	56.4	59.2
15	2,500	70.7	66.8	68.4	66.7	62.0	62.7
0	500	75.1	71.8	68.9	59.5	56.1	61.2
0	1,000	89.0	84.8	78.8	82.2	69.2	67.0
0	2,500	98.4	96.2	91.7	97.6	91.3	78.0
Combination effects							
Overall	Overall	78.4	76.0	75.0	70.8	65.4	68.3

A.3: Discrimination of relevant versus irrelevant variables for several techniques to estimate variable importance. *(continued)*

Max correlation	No. observations	Negation	ANOVA	Permutation	SHAP	Gain	Permutation
30	500	65.1	64.0	63.4	55.6	50.0	59.5
30	1,000	67.5	65.6	65.3	61.7	55.8	61.0
30	2,500	70.1	67.5	68.6	66.0	62.7	64.7
15	500	70.3	68.1	66.7	58.7	52.4	62.2
15	1,000	74.8	71.1	71.5	66.5	59.8	64.4
15	2,500	78.8	74.9	77.0	73.0	68.9	70.2
0	500	83.8	81.1	76.5	66.8	62.1	67.5
0	1,000	95.6	92.6	87.9	89.6	78.9	76.3
0	2,500	99.8	99.3	97.8	99.6	97.8	88.9
Main effects							
Overall	Overall	91.0	88.9	88.7	85.0	82.7	83.0
30	500	79.4	77.0	75.5	70.1	66.3	70.7
30	1,000	83.6	80.6	80.6	77.0	73.9	74.6
30	2,500	86.9	83.8	85.4	82.2	80.9	79.1
15	500	86.3	83.3	81.7	74.6	70.4	74.4
15	1,000	91.0	87.9	88.2	84.7	81.4	80.5
15	2,500	94.6	91.7	93.7	90.2	88.7	86.7
0	500	97.7	96.2	94.0	86.6	84.5	85.9
0	1,000	99.9	99.8	99.3	99.5	97.9	95.0
0	2,500	100.0	100.0	100.0	100.0	100.0	99.8

References

Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *The Journal of Machine Learning Research*, 18(1):2653–2688, 2017.

Paul Blanche, Jean-François Dartigues, and Hélène Jacqmin-Gadda. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in medicine*, 32(30):5381–5397, 2013.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Routledge, 2017.

Yifan Cui, Ruoqing Zhu, Mai Zhou, and Michael Kosorok. Consistency of survival tree and forest models: splitting bias and correction. *arXiv preprint arXiv:1707.09631*, 2017.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006.

Ben Goodrich, Jonah Gabry, Imad Ali, and Sam Brilleman. rstanarm: Bayesian applied regression modeling via Stan., 2022. URL `https://mc-stan.org/rstanarm/`. R package version 2.21.3.

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545, 1999. URL `https://doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18%3C2529::AID-SIM274%3E3.0.CO;2-5`.

Frank E. Harrell, Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. Evaluating the Yield of Medical Tests. *JAMA*, 247(18):2543–2546, 05 1982. ISSN 0098-7484. doi: 10.1001/jama.1982.03320430047030. URL `https://doi.org/10.1001/jama.1982.03320430047030`.

David Heath, Simon Kasif, and Steven Salzberg. Induction of oblique decision trees. In *IJCAI*, volume 1993, pages 1002–1007. Citeseer, 1993.

Torsten Hothorn, Berthold Lausen, Axel Benner, and Martin Radespiel-Tröger. Bagging survival trees. *Statistics in medicine*, 23(1):77–91, 2004.

Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15 (3):651–674, 2006.

Hemant Ishwaran and Udaya B Kogalur. Consistency of random survival forests. *Statistics & probability letters*, 80(13-14):1056–1064, 2010.

Hemant Ishwaran and Min Lu. Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in medicine*, 38(4):558–582, 2019.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The Annals of Applied Statistics*, pages 841–860, 2008.

Byron C Jaeger, D Leann Long, Dustin M Long, Mario Sims, Jeff M Szychowski, Yuan-I Min, Leslie A Mcclure, George Howard, and Noah Simon. Oblique random survival forests. *The Annals of Applied Statistics*, 13(3):1847–1883, 2019.

Michael W Kattan and Thomas A Gerds. The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. *Diagnostic and prognostic research*, 2(1):1–7, 2018.

Rakesh Katuwal, Ponnuthurai Nagaratnam Suganthan, and Le Zhang. Heterogeneous oblique random forest. *Pattern Recognition*, 99:107078, 2020.

Max Kuhn and Hadley Wickham. *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.*, 2020. URL `https://www.tidymodels.org`.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

Bjoern H Menze, B Michael Kelm, Daniel N Splitthoff, Ullrich Koethe, and Fred A Hamprecht. On oblique random forests. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 453–469. Springer, 2011.

Karel GM Moons, Andre Pascal Kengne, Diederick E Grobbee, Patrick Royston, Yvonne Vergouwe, Douglas G Altman, and Mark Woodward. Risk prediction models: II. external validation, model updating, and impact assessment. *Heart*, 98(9):691–698, 2012a.

Karel GM Moons, Andre Pascal Kengne, Mark Woodward, Patrick Royston, Yvonne Vergouwe, Douglas G Altman, and Diederick E Grobbee. Risk prediction models: I. development, internal validation, and assessing the incremental value of a new (bio) marker. *Heart*, 98(9):683–690, 2012b.

Sreerama K Murthy, Simon Kasif, and Steven Salzberg. A system for induction of oblique decision trees. *Journal of Artificial Intelligence Research*, 2:1–32, 1994.

Nitesh Poona, Adriaan Van Niekerk, and Riyad Ismail. Investigating the utility of oblique tree-based ensembles for the classification of hyperspectral data. *Sensors*, 16(11):1918, 2016.

Xueheng Qiu, Le Zhang, Ponnuthurai Nagaratnam Suganthan, and Gehan AJ Amaratunga. Oblique random forest ensemble via least square estimation for time series forecasting. *Information Sciences*, 420:249–262, 2017.

Tom Rainforth and Frank Wood. Canonical correlation forests. *arXiv preprint arXiv:1507.05444*, 2015.

Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25, 2007.

Terry Therneau. Survival package source code documentation, April 2022. URL `https://github.com/therneau/survival/blob/5440691d44abea537b08aeb60153a31654d66a9b/noweb`. original-date: 2016-04-28.

Tyler M Tomita, James Browne, Cencheng Shen, Jaewon Chung, Jesse L Patsolic, Benjamin Falk, Carey E Priebe, Jason Yim, Randal Burns, Mauro Maggioni, et al. Sparse projection oblique randomer forests. *Journal of machine learning research*, 21(104), 2020.

Hong Wang and Gang Li. A selective review on random survival forests for high dimensional data. *Quantitative bio-science*, 36(2):85, 2017.

Hong Wang and Lifeng Zhou. Random survival forest with space extensions for censored data. *Artificial intelligence in medicine*, 79:52–61, 2017.

Marvin N. Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *Journal of Statistical Software*, 77(1):1–17, 2017. doi: 10.18637/jss.v077.i01. URL `https://www.jstatsoft.org/index.php/jss/article/view/v077i01`.

Le Zhang and Ponnuthurai N Suganthan. Oblique decision tree ensemble via multisurface proximal support vector machine. *IEEE transactions on cybernetics*, 45(10):2165–2176, 2014.

Lifeng Zhou, Hong Wang, and Qingsong Xu. Random rotation survival forest for high dimensional censored data. *SpringerPlus*, 5(1):1–10, 2016.

Ruoqing Zhu. *Tree-based Methods for Survival Analysis and High-dimensional Data*. PhD thesis, The University of North Carolina at Chapel Hill, 2013.

Ruoqing Zhu, Donglin Zeng, and Michael R Kosorok. Reinforcement learning trees. *Journal of the American Statistical Association*, 110(512):1770–1784, 2015.